МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ

ГОСУДАРСТВЕННОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ

ВЫСШЕГО ПРОФЕССИОНАЛЬНОГО ОБРАЗОВАНИЯ

«УЛЬЯНОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»

ИНСТИТУТ МЕЖДУНАРОДНЫХ ОТНОШЕНИЙ

КАФЕДРА АНГЛИЙСКОГО ЯЗЫКА ДЛЯ ПРОФЕССИОНАЛЬНОЙ ДЕЯТЕЛЬНОСТИ

Э.В. Егорова, Н.А. Крашенинникова, О.И. Осетрова

Методическое пособие по домашнему чтению для студентов неязыковых специальностей вузов

Ульяновск
2014

Предисловие.

Данное учебное пособие представляет собой методическое пособие по домашнему чтению для студентов неязыковых специальностей вузов. Пособие составлено в соответствии с новой программой курса английского языка, предусмотренного модульной системой образования. Пособие включает рекомендации по выбору литературы, непосредственно алгоритм работы над домашним чтением и анализ основных ошибок, возникающих у студентов в ходе работы.

Целью данного учебного пособия является совершенствование языковых компетенций в рамках самостоятельной работы над домашним чтением, в ходе которого формируется устойчивый интерес к чтению как средству познания других культур, а также знания о конкретных социокультурных реалиях. Отличительной особенностью данного пособия является четкий пошаговый алгоритм работы с домашним чтением, наличие рекомендаций по тому, что необходимо и нельзя делать в процессе выполнения работы. Пособие ориентировано на формирование профессионально значимых компетенций, необходимых для чтения литературы по изучаемой специальности с целью извлечения информации и составления аннотации прочитанного.

Пособие состоит из 6 разделов и приложений, в которых наглядно иллюстрируются все основные этапы домашнего чтения. Четко следуя предлагаемым авторами рекомендациям, студенты смогут без труда выполнить, оформить и успешно сдать отчет по домашнему чтению.

Авторы надеются, что данное методическое пособие будет полезно всем студентам, обучающимся на неязыковых специальностях.

I. Внеаудиторное (домашнее) чтение как вид учебной деятельности по иностранному языку.

В настоящее время концепция обучения иностранному языку предусматривает формирование коммуникативной компетенции учащихся. Речевая компетенция предполагает развитие коммуникативных умений в четырех основных видах речевой деятельности: говорении, аудировании, чтении и письме. Известно, что чтение оригинальной литературы (как художественной, так и специализированной) на изучаемом языке значительно обогащает словарный запас, знакомит с культурой и литературой страны изучаемого языка, развивает аналитическое мышление, а также способствует развитию устной речи.

Однако, исходя из современной практики, в рамках классно-урочной системы присутствует в основном «интенсивное» чтение, то есть чтение небольших по объему текстов. Получается, что за многие годы изучения иностранного языка учащиеся могут не прочитать ни одной книги на иностранном языке. Это связано, с одной стороны, с нежеланием читать книги вообще (как на родном, так и на иностранном языке), а с другой стороны, с «боязнью» непривычно большого объема текста, незнакомых лексических и грамматических структур, языковых реалий и т.д. [4, 9].

В Федеральном государственном образовательном стандарте прописана необходимость обеспечить «сформированность устойчивого интереса к чтению как средству познания других культур». В то же время современная концепция предмета «иностранный язык» связана с осуществлением на старшей ступени профильного обучения. Это означает, что иностранный язык должен быть связан с избранным профилем, например психологическим, историческим, экологическим, культурологическим и т.д.

Кроме того, согласно новым образовательным стандартам, достаточно большое количество часов выделяется на самостоятельную работу студентов. А одним из основных и наиболее доступных видов самостоятельной работы является, на наш взгляд, домашнее чтение. Следовательно, курс домашнего

чтения на английском языке должен дополнять основной курс английского языка, выполняя ряд функций.

Первостепенные задачи.

1. *Извлечение информации из текста* в том объеме, который необходим для решения конкретной речевой задачи, используя определенные технологии чтения [3].

2. *Формирование интереса к чтению на английском языке.* Работая с книгами для чтения разных жанров под руководством преподавателя, учащиеся приобретут большую уверенность в своих силах, научатся выбирать подходящие по уровню книги, преодолевать языковые трудности, почувствуют вкус к литературе.

3. *Углубление знаний в области культуры стран изучаемого языка.* Широкий выбор книг для домашнего чтения позволит учащимся познакомиться с лучшими образцами современных произведений зарубежных авторов, узнать больше о традициях англоязычных стран, их истории и образе жизни. Домашнее чтение помогает развивать и совершенствовать общие представления о мире. Зачастую студенты обладают достаточно ограниченными представлениями о мире, в котором они проживают. Домашнее чтение помогает расширить общий кругозор обучающихся.

4. *Обучение литературному анализу.* При выполнении заданий по прочитанным книгам, изложении основного содержания, составлении характеристик героев, совершенствовании своих знаний в области литературных приемов, у студентов будут продолжать формироваться навыки анализа литературных произведений, база которых, согласно ФГОС, должна быть заложена в рамках средней школы.

5. *Формирование умения «использовать иностранный язык как средство для получения информации из иноязычных источников в образовательных целях».* Использование научно-популярной и другой нехудожественной литературы для углубленного домашнего чтения позволит

студентам овладеть иностранным языком в сфере профессиональной коммуникации, получить и оценить информацию из современных зарубежных источников и использовать её в своей научной деятельности.

Вспомогательные задачи.

1. *Расширение словарного запаса* учащихся за счет лексических единиц текстов книг и, что особенно важно, устойчивых словосочетаний. Использование домашнего чтения позволит учащимся оптимизировать процесс усвоения языкового и речевого материала. Лексические единицы нельзя выучить, увидев их единожды. Домашнее чтение позволяет многократно столкнуться со словами и выражениями в контексте, способствуя их более прочному запоминанию. Кроме того представление новых слов в контексте способствует развитию языковой догадки и чутья. Британский лингвист Michael Hoey также подчеркивает значительный эффект многократного представления новых лексических единиц в контексте [7].

2. *Дальнейшее развитие навыков* не только чтения, но и *аудирования*, так как многие книги для чтения имеют аудиосопровождение.

3. *Дальнейшее развитие устной монологической и диалогической речи* при помощи пересказов разных видов, обсуждения прочитанного, дискуссий, составления презентаций и др.

4. Формирование навыков творческого письма через выполнение специальных заданий. Такие задания помогут дальнейшему формированию и отработке всех языковых и речевых навыков [4]. В принципе, любое чтение помогает совершенствовать навыки письма. Существует совершенно очевидная взаимосвязь между чтением и письмом: чем больше мы читаем, тем лучше мы пишем. Хотя процесс этой взаимосвязи еще не до конца изучен [8], но сам факт того, что это происходит, задокументирован [6]. С точки зрения здравого смысла, чем больше языкового материала мы прорабатываем, чем чаще мы читаем, тем вероятнее, что наш механизм

языкового восприятия будет помогать воспроизводить увиденные нами языковые единицы на письме или в речи по мере необходимости.

Наряду с организацией домашнего чтения большое значение имеет содержательная сторона учебных материалов, предназначенных для чтения [2]. Ведь важно, чтобы студенты не просто читали большое количество текстов на иностранном языке, но чтобы они получали от этого удовольствие. Поэтому студентам рекомендуется выбирать книгу для чтения в зависимости от их интересов, уровня подготовки, актуальности материала, тематической близости предмета изложения к жизненному опыту и т.д. [5]. Исследователи говорят, что только неподдельный интерес поможет сконцентрироваться на произведении и взять из него максимум полезной информации. В настоящее время доступно огромное количество книг для чтения, имеющих познавательное или развлекательное содержание («Reading for Pleasure»), причем зачастую книги могут быть распределены по уровням владения языком. Необходимо помнить, что если студент находит книгу слишком сложной, или она ему не нравится, он всегда может взять другую. В то же время, слишком простая книга может демотивировать студента из-за заблуждения, что она ничему не сможет его научить [9].

Более того, в своей книге «Вас невозможно научить иностранному языку» Н.Ф. Замяткин говорит, что отдавать предпочтение нужно целым книгам, а не отдельным фрагментам [1]. Дело в том, что работая с длинным произведением, читатель приспосабливается к стилю автора, его логике изложения материала. Это, в свою очередь, помогает лучше распознавать содержание и запоминать лексические и грамматические конструкции.

При контроле внеаудиторного (домашнего) чтения аутентичной литературы по специальности, который проводится на индивидуальных консультациях, проверяется наличие умений в разных видах речевой деятельности:

- бегло читать оригинальную литературу в соответствующей отрасли знания;

- оформлять извлеченную из текста информацию в виде перевода или резюме;
- делать сообщения и доклады на иностранном языке на темы, затрагиваемые в прочитанной литературе;
- вести беседу по прочитанному материалу.

Формирование вышеперечисленных умений требует самостоятельной работы со словарем, внеаудиторной работы по оформлению перевода специальной литературы, самостоятельной работы по составлению логически последовательного, полного и мотивированного монологического высказывания (в виде сообщения или доклада). Составление словаря терминов, входящих в терминологическую систему конкретного подъязыка, должно производиться также самостоятельно на протяжении всего курса обучения.

II. Требования к отчету по домашнему чтению.

Студенты всех специальностей ФГБОУ ВПО «Ульяновский государственный университет» в обязательном порядке изучают иностранный язык в объеме, предусмотренном программой и состоящем из часов, отведенных на аудиторные занятия и самостоятельную работу студентов. Именно последний вид работы предполагает обязательное выполнение заданий по домашнему (внеаудиторному) чтению в течение семестра. В соответствии с нормативными требованиями кафедры английского языка гуманитарных специальностей **отчет по домашнему чтению** за семестр включает в себя:

1. **титульный лист**, который должен содержать название учебного заведения, название кафедры, тему работы, фамилию, инициалы студента, номер группы, фамилию, инициалы, ученые и академические звания научного руководителя, название города, а также год написания работы (см. Приложение 1);

2. **письменный перевод** научного текста по специальности (Translation) объемом 45 тысяч печатных знаков (далее «т.п.з.»), включая пробелы и знаки препинания;

Примечание.

При написании перевода необходимо руководствоваться следующими требованиями. Объем письменного перевода должен составлять не менее 45 т.п.з. оригинального текста (на английском языке). Оформляется на отдельных стандартных листах формата А4:

- *14 кегль, шрифт Times New Roman,*
- *интервал в 1,5 строки,*
- *выравнивание по ширине,*
- *левое поле – 30 мм, верхнее – 20 мм, правое – 20 мм, нижнее – 25 мм.*

*Первая часть отчета (Translation) представляет собой **параллельный перевод**, выполненный в виде таблицы, в которой абзацы*

английского и русского текстов ***четко соответствуют друг другу*** (см. Приложение 2). *Печатать следует на одной стороне листа. Все страницы нумеруются. Первой страницей считается титульный лист, на ней номер не ставится, на следующей странице проставляется цифра 2 и так далее. Порядковый номер печатается в правом нижнем углу страницы. Работы, не соответствующие требованиям оформления, к проверке не принимаются и подлежат доработке.*

3. краткий (1,5 – 2,5 т.п.з.) **пересказ** прочитанного на английском языке (Summary) в соответствии с правилами пересказа, выполненный на отдельном листе с сохранением сплошной нумерации страниц (см. Приложение 3);

4. **грамматическое задание** (Grammar Task) на согласованную с преподавателем тему: 10 предложений из исходного текста, содержащих соответствующие грамматические единицы, а также перевод этих предложений на русский язык (выполняется на отдельном листе с сохранением сплошной нумерации страниц); соответствующая грамматическая конструкция выделяется в английском предложении подчеркиванием (см. Приложение 4);

5. список **ключевых слов** (Key Words) из текста (100 единиц), снабженных транскрипцией и переводом на русский язык (см. Приложение 5) (выполняется на отдельном листе с сохранением сплошной нумерации страниц); ключевые слова необходимо выучить наизусть и сдать преподавателю; в список ключевых слов нельзя включать активную лексику (слова, которые изучаются на аудиторных занятиях с преподавателем во время семестра);

6. **постраничный словарь** всех лексических единиц, вызвавших затруднения при переводе с указанием на расположение в тексте (страница, номер абзаца и / или строки). Постраничный словарь выполняется в произвольной форме, обязателен для предъявления на всех этапах контроля, однако вместе с отчетом не сдается и не оценивается.

Примечание.

В соответствии с принципом дифференцированного обучения преподаватель вправе применять индивидуальный подход и изменять представленные нормативы в ту или иную сторону в интересах повышения эффективности учебного процесса.

Предусмотрены два вида контроля выполнения домашнего чтения студентами: промежуточный и итоговый. **Промежуточный** проводится регулярно, **1 раз в 2-4 недели** в течение семестра, а также непосредственно перед **внутрисеместровой аттестацией**. Преподаватель предлагает студентам представить предварительные результаты работы: элементы постраничного словаря, устный или письменный перевод готового фрагмента, выученные лексические единицы из списка ключевых слов. Это делается для того, чтобы стимулировать студента планировать свою работу заблаговременно, грамотно распределяя силы и время, иначе спешка и большой объем задания могут сказаться на качестве работы. Кроме того, подсказки и советы преподавателя позволят студентам исправить возможные недочеты в работе и не совершать их в дальнейшем.

Примечание.

*Преподаватель имеет право **не аттестовать** студента за неудовлетворительное текущее выполнение заданий по домашнему чтению.*

Итоговый контроль проводится для всей группы в специально отведенное преподавателем время **за 3-4 недели до конца семестра** и предполагает:

1) проверку преподавателем качества письменного перевода, краткого пересказа и грамматического задания;

2) выполняемые студентом чтение вслух и устный перевод выбранного преподавателем фрагмента текста;

3) устное воспроизведение студентом английских эквивалентов русских лексических единиц из предварительно подготовленного списка ключевых слов.

Примечание.

Общая неудовлетворительная оценка за домашнее чтение как неотъемлемый компонент программы по иностранному языку ведет к ***снижению экзаменационной оценки на 1 балл****. В день экзамена / зачета контроль домашнего чтения не производится.*

III. Требования к тексту оригинала и словарям.

Одним из важнейших условий для успешного выполнения заданий по домашнему чтению является правильный выбор источника, который должен удовлетворять ряду требований.

Во-первых, это должен быть **аутентичный текст**, т.е. изначально созданный на английском языке носителем этого языка. Именно по этой причине некоторые классические тексты не годятся для выполнения перевода по домашнему чтению. Так, большинство средневековых авторов, например, Беда Достопочтенный, Фрэнсис Бэкон и другие писали на латыни. А такие авторы, как Зигмунд Фрейд, Эмиль Дюркгейм, Огюст Конт и другие авторы – на родных языках (немецкий, французский и т.п.). Как правило, английские тексты этих авторов являются переводами с оригинала.

Во-вторых, не следует брать произведения художественной литературы, т.к. именно те качества, которые нас в ней привлекают (образность, стилевое разнообразие, наполненность различными аллюзиями), сильно усложняют работу по переводу. Кроме того, целью обучения в вузе является, прежде всего, расширение профессиональных компетенций. Поэтому тексты нужно подбирать научные или хотя бы научно-популярные. Для некоторых специальностей могут подойти тексты, ориентированные на практическую деятельность (социальная работа, таможенное дело, лесное хозяйство). В любом случае рекомендуется в каждом отдельном случае консультироваться с преподавателем.

В-третьих, тексты для домашнего чтения важно подбирать в соответствии со своими научными и профессиональными интересами. Домашнее чтение ни в коем случае не должно превращаться в пытку. Более того, при определенном подходе этот вид деятельности со временем может стать источником вдохновения, новых знаний, идей, да и просто хобби. Бывает, студенты ссылаются на тексты домашнего чтения при выполнении курсовых и дипломных работ по специальности. Кто-то использует приобретенные навыки в работе уже по окончании университета, читая

самые свежие материалы по своей профессии в оригинале, до того, как с ними успели ознакомиться в переводе (не всегда адекватном) их конкуренты. Нельзя не упомянуть о том, что при обучении в аспирантуре предусмотрен экзамен кандидатского минимума по иностранному языку, предполагающий, помимо прочего, работу с оригинальным текстом по специальности объемом 500 т.п.з. Поэтому умение выбрать для чтения интересную и нужную книгу, а также правильно оформить результаты своего труда – навык отнюдь в жизни не лишний, и работа над его формированием должна начаться на первых неделях обучения в университете.

Для домашнего чтения подойдут как печатные издания, так и их электронные варианты. Что касается первых, то отдел литературы на иностранных языках библиотеки УлГУ обладает хорошим выбором как классической, так и современной научной литературы. Подобрать книгу на интересующую тему можно при помощи каталога, подобно литературе на русском языке. Можно также воспользоваться списком на сайте кафедры АЯПД. Минус такого выбора состоит в том, что отрывок, предназначенный для перевода, придется сканировать и распознавать, чтобы позже поместить в таблицу параллельного перевода. Подходящий текст в электронном виде обработать проще, но бывает сложнее подобрать, т.к. не к кому обратиться за помощью в случае затруднений при выборе текста (см. Приложение 6).

Определившись с исходным материалом для перевода, следует подумать и об инструменте его выполнения. Речь идет о словарях, которые тоже, в свою очередь, доступны и в книжном формате, и в электронном. Здесь все зависит от индивидуальных предпочтений студента. Прежде всего, хотелось бы процитировать уже упоминавшегося выше Н.Ф. Замяткина: «При покупке словарей есть одни интересные «грабли», на которые очень многие наступают. Сначала они покупают самый маленький словарик. Через достаточно короткое время обязательно обнаруживается, что такого словаря недостаточно. Покупается второй словарь – чуть больше размером. Потом третий и так далее – вплоть до приобретения самого толстого словаря. Таким

образом, у вас дома лежит, обрастая пылью, совершенно бесполезная коллекция разнокалиберных словарей – за исключением действительно необходимого для вас последнего словаря, именно с которого вам было нужно и начинать ваше словарное приключение. Так что сразу начинайте с конца и раз и навсегда приобретайте самый большой словарь, не занимаясь коллекционированием макулатуры». Действительно, в миниатюрных словарях содержится ограниченное количество слов и значений, что обязательно негативно повлияет на понимание текста и качество перевода. В качестве достойного варианта авторы могут порекомендовать, например, «Англо-русский словарь» В.К. Мюллера, или электронный словарь www.multitran.ru. Подойдут также словари серии ABBYY Lingvo. А вот чего совсем не стоит делать, так это пользоваться электронными переводчиками. Впрочем, об этом будет сказано ниже.

IV. Методика работы над отчетом по домашнему чтению.

Резюмируя изложенные в предыдущих разделах требования и правила, предлагаем примерный алгоритм работы над домашним чтением.

Итак, преподаватель проинформировал Вас об объеме и сроках сдачи отчета по домашнему чтению. Студентам второго и последующих курсов, как правило, проще определиться со своими профессиональными интересами и выбрать соответствующий материал для работы. Однако хочется верить, что и первокурсники уже имеют некоторое представление о специальности, по которой они будут проходить обучение в вузе в течение ближайших четырех или пяти лет.

1. Выбрав интересующую Вас область (например, для студентов-историков это может быть война за независимость США, для будущих социальных работников – геронтология и т.п.), определитесь с типом источника. Если Вы предпочитаете печатные книги, то Вас ждет Отдел литературы на иностранных языках научной библиотеки УлГУ (второй корпус, второй этаж, ауд. 22). Здесь по праву гордятся прекрасной подборкой современной научной литературы из Оксфордского и Кембриджского фондов. Если у Вас уже есть читательский билет, можно начать работу над домашним чтением с отдела каталогов (на это потребуется 1,5 – 2 часа):

1.1. найдите в систематическом каталоге ящичек с нужной Вам темой;

1.2. ближе к концу темы найдите карточку с надписью «книги на иностранных языках»;

1.3. с помощью предварительно взятого с собой словаря переведите названия найденных книг и отберите несколько из них, которые, как Вам кажется, могли бы Вас заинтересовать;

1.4. если таковых не окажется, повторите процедуру с другими интересными Вам темами, пока не подберете 3-5 вариантов;

1.5. спишите выходные данные и библиотечный шифр этих книг, как Вас учили на занятиях по библиографии (в случае затруднений обратитесь к сотрудникам библиотеки);

1.6. идите в читальный зал отдела литературы на иностранных языках и попросите выдать всю литературу из получившегося списка (часть книг может быть на руках);

Примечание.

Можно начать поиск книги непосредственно в отделе литературы на иностранных языках, воспользовавшись электронным каталогом и поиском по ключевым словам (в этом случае будет целесообразно заранее узнать перевод интересующих Вас терминов или ключевых слов на английский язык, например, medieval markets, criminal law, child psychology, ecosystems, и т.д.).

1.7. внимательно просмотрите все книги, начиная с оглавления, при этом постарайтесь получить хотя бы приблизительное представление об их содержании;

1.8. если есть возможность, пробуйте узнать что-нибудь об авторе или даже о самой книге (воспользуйтесь русскоязычным сектором интернета);

1.9. обратите внимание, что таблицы, схемы, диаграммы и картинки не будут учитываться при подсчете объема домашнего чтения;

1.10. обратите внимание на четкость и удобочитаемость шрифта, а также на другие факторы, которые могут помешать адекватному распознаванию текста (а Вам это обязательно предстоит сделать);

1.11. верните книги, которые Вам не подошли, сотруднику библиотеки и оформите выбранную книгу в соответствии с библиотечными правилами;

Примечание.

Помните, что библиотека обслуживает огромное количество студентов, сотрудников и преподавателей. Работая с печатным изданием, не пачкайте его, сведите к минимуму карандашные пометки, не вырывайте страницы. После Вас книга может потребоваться другому человеку, например, декану Вашего факультета. Не стоит создавать себе лишние сложности, ведь узнать, кто брал книгу до Вас, очень просто.

1.12. произведите сканирование требуемого количества страниц книги, а также распознавание и импортирование результатов в Word своими силами или воспользовавшись услугами библиотеки.

Примечание.

Приблизительный подсчет требуемого объема текста производится следующим образом:

- *подсчитывается среднее количество печатных знаков на странице (нужно сосчитать количество всех знаков на одной строке, включая пробелы и знаки препинания, и умножить его на количество строк);*
- *общий нормативный объем (45 т.п.з.) делится на количество печатных знаков на странице (см. предыдущий пункт);*
- *в итоге получается количество страниц, которые следует сканировать (постарайтесь, чтобы текст представлял собой законченный в смысловом плане отрывок).*

Если Вам проще работать сразу с электронным документом, воспользуйтесь списком интернет-ресурсов (Приложение 6) или самостоятельно поищите в интернете текст для перевода. Советуем пользоваться англоязычными поисковыми системами, например, Google. Поиск желательно проводить по ключевым словам (см. Примечание к п. 1.6). Не обязательно искать именно книгу, попробуйте подобрать несколько статей на интересующую Вас тему. С материалами в электронном виде тоже не мешает предварительно ознакомиться, прежде чем скачивать или распечатывать.

2. Работа с текстом.

2.1. Предварительное оформление:

2.1.1. для подсчета количества печатных знаков в документе следует воспользоваться функцией «статистика», выделив нужный фрагмент в

открытом документе Word. Если Вы не выделите фрагмент, программа произведет подсчет количества знаков во всем документе;

2.1.2. следует учитывать все знаки, **включая пробелы и знаки препинания**;

2.1.3. схемы, таблицы, графики, диаграммы и т.п. **не учитываются**;

2.1.4. в документе Word создайте **таблицу из двух колонок**: первая будет содержать текст оригинала (на английском языке), вторая – Ваш перевод (на русском языке);

2.1.5. абзацы оригинала и перевода должны **совпадать**, для этого после каждого абзаца открывайте в таблице новую строку (для удобства можно открывать новую строку после каждого предложения);

2.1.6. помните о требованиях к оформлению (см. Раздел III).

2.2. Работа над переводом:

2.2.1. если Вы предпочитаете пользоваться электронным словарем (например, словари Яндекс, Multitran), то целесообразно свернуть документ Word в окно и скорректировать его размер, то же самое рекомендуется сделать с открытым окном словаря;

2.2.1. если Вы пользуетесь англо-русским словарем в книжном формате, заранее сделайте закладки из тонких полос бумаги на каждой букве английского алфавита (на закладке четко пишется соответствующая буква, закладки располагаются таким образом, чтобы все они были видны одновременно), что может значительно облегчить и ускорить поиск нужного слова, особенно если Вы не уверены в своем знании порядка букв английского алфавита;

2.2.2. приготовьте тетрадь на 24-48 листов для создания своего индивидуального постраничного словаря;

2.2.3. подпишите тетрадь, на первой странице напишите информацию, которая пригодится для оформления отчета по домашнему чтению: название текста и фамилию автора, место и год издания, как они написаны в книге, а

также номера страниц, которые собираетесь переводить (например, стр. 5-37).

2.3. Теперь можно приступить к работе над **постраничным словарем:**

2.3.1. разверните тетрадь и расчертите разворот на 4 графы: на первой странице – английское слово, транскрипция и русский эквивалент (перевод), а вторую страницу отведите под примечания (заносите туда детали из словарной статьи, которые покажутся Вам важными: сочетаемость с другими словами, устойчивые выражения, дополнительные значения – это может сэкономить Вам время при переводе);

2.3.2. начните заполнение словаря с номера и названия той главы книги, которую переводите (напишите его так, как оно дано в книге, выделите цветом или подчеркиванием);

2.3.2. на следующей строке напишите номер страницы, с которой начинаете переводить (его также следует выделить);

2.3.3. отметьте номер абзаца / строки (последний пункт можно опустить студентам с уровнем владения английским языком не ниже Upper Intermediate ввиду богатого словарного запаса);

2.3.4. каждый раз, переходя к переводу следующей страницы, четко указывайте ее номер (это поможет Вам лучше ориентироваться в своем постраничном словаре во время проведения устного контроля);

2.3.5. прочитайте первое предложение без помощи словаря, попытайтесь хотя бы в общих чертах представить, о чем идет речь;

2.3.6. прочитайте предложение еще раз, выписывая в первую колонку **все** слова, который вызвали у Вас затруднения, даже если они кажутся Вам маленькими и незначительными, ведь именно они зачастую полностью меняют смысл «больших» слов;

Примечание:

Не пишите слова в каждой строчке – Вам может понадобиться место для тех слов, в значении которых Вы сначала были уверены, а потом засомневались.

2.3.7. прежде чем начинать работу со словарем, попробуйте по внешним, формальным признакам (суффиксы, окончания, сочетание с другими словами), определить, к каким частям речи принадлежат выписанные слова, а также какими членами предложения они являются;

Примечание:

Имейте ввиду, что

- *подлежащее (чаще всего оно выражено именем существительным или местоимением) обычно расположено в начале предложения перед сказуемым, которое может состоять из нескольких частей (в основном это вспомогательные глаголы be, have, do в разных формах);*
- *после сказуемого следует прямое дополнение, затем косвенное (с предлогом), после них – обстоятельства места и времени (дополнение, как правило, бывает выражено существительным или местоимением, а обстоятельства – существительным с предлогом или наречием);*
- *определение, как правило, находится перед определяемым словом, причем оно далеко не всегда выражено прилагательным или причастием, зачастую эту функцию выполняет существительное (например, door knob – дверная ручка).*

2.3.8. найдите выписанное слово в словаре, прочитайте вслух при помощи транскрипции, при необходимости перепишите транскрипцию во вторую колонку;

Примечание:

Если Вы испытываете сложности с чтением транскрипции, наберите нужное слово в поисковике любого электронного словаря и кликните на значке с изображением динамика, чтобы прослушать, как оно звучит. Повторите за диктором несколько раз, пока не будете уверены, что Ваше воспроизведение похоже на оригинал.

2.3.9. внимательно полностью прочитайте соответствующую словарную статью, обращая внимание на принадлежность слова к определенной части речи (например, слово *question* может оказаться как существительным, так и глаголом), а также – на значение именно в этом контексте (например, слово *gate* в одном контексте может переводиться как *ворота*, а в другом – как *денежный сбор*);

Примечание:

Не забывайте об изменении значения слов из-за сочетания с другими словами:

- *многие глаголы очень зависят от предлогов (look at – смотреть, look after – заботиться, присматривать, look for – искать);*
- *есть много устойчивых выражений, значение которых не соответствует переводу отдельных входящих в них слов (например, to be on the air следует переводить не «быть на воздухе», а «выступать по радио»), поэтому обязательно еще раз просмотрите выписанные слова и, при необходимости, само предложение – возможно, Вы пытаетесь «изобрести велосипед», хотя все необходимое за Вас уже сделали составители словаря;*
- *иногда все слова в предложении кажутся знакомыми, но смысл, тем не менее, ускользает. Основных причин может быть две: наличие неизвестных Вам значений у «знакомых» слов (в этом случае не поленитесь хотя бы на скорую руку просмотреть все значения тех слов, которые кажутся Вам знакомыми, и Вы удивитесь обилию новых для Вас смыслов) и сложность грамматической конструкции. В последнем случае поможет только подробный грамматический разбор по членам предложения и грамматический справочник под рукой;*
- *именно проблемы, изложенные авторами в этом примечании, являются главной слабостью большинства электронных переводчиков: они не умеют делать правильные умозаключения и просто предлагают наиболее часто употребляемые лексические и грамматические модели.*

Результат можно было бы назвать смешным, если бы это не было так грустно! Поверьте, даже если Ваши познания в английском языке оставляют желать лучшего, Вы вполне в состоянии сделать перевод более грамотно, чем машина, нужно только немного усилий, терпения и времени.

2.3.10. выпишите в третью колонку Вашего постраничного словаря то значение, которое кажется наиболее подходящим, а в четвертую – другие возможные варианты, сочетания;

2.3.11. выписывая слово в **определенной грамматической форме**, например, существительное во множественном числе или глагол во второй форме, Вы должны **сохранить эту форму**, когда записываете русский эквивалент в третью колонку: *takes* – не «брать», а «берет», *maps* – не «карта», а «карты», *caught* – не ловить, а «ловил» или «пойманный».

2.4. Подробно разобрав предложение при помощи постраничного словаря, поняв смысл, приступайте к его **переводу на русский язык:**

2.4.1. некоторые предпочитают сначала выполнять черновой письменный вариант, а затем набирать его на компьютере. У такого подхода есть свои плюсы: скорость перевода не зависит от Вашей скорости печати, можно видеть не только конечный вариант перевода, но и предварительные (возможно, как раз один из них и окажется более точным), не нужен доступ к компьютеру. Однако мы все-таки рекомендуем сразу набирать перевод на компьютере. Во-первых, не нужно вертеть головой в поисках текста оригинала, т.к. он всегда перед глазами. Во-вторых, скорость печати – дело наживное и чем больше Вы печатаете, тем она будет выше. В-третьих, откладывая набор текста на потом, Вы рискуете задержать сдачу отчета, а преподаватель не будет принимать работу-полуфабрикат;

2.4.2. попробуйте передать на русском языке то, что автор сказал на английском – это и есть перевод – и наберите результат Ваших усилий в правой колонке таблицы документа Word, предназначенной для перевода (в

левой колонке, как Вы помните из п.п. 2.1.4. и 2.1.5., расположен оригинальный текст);

2.4.3. помните, что Вы пишете перевод на родном языке, поэтому **не забывайте согласовывать грамматические формы** (время, лицо, число, падеж)**;**

2.4.4. выполнив перевод всех предложений абзаца, просмотрите их еще раз, убедитесь, что Вы понимаете, о чем идет речь, и что в Вашем переводе нет противоречий;

2.4.5. проделайте ту же процедуру (п.2.4.4.) по окончании работы над всем переводом (зачастую ближе к окончанию работы содержание становится яснее). Возможно, Вам стоит изменить что-то в первых абзацах.

Примечание

Если Вы старательно и своевременно работаете над своим заданием по домашнему чтению, то нет необходимости обращаться за посторонней помощью, которую часто предлагают желающие подзаработать «специалисты по иностранным языкам». Некоторые из них даже предлагают сделать постраничный словарь за отдельную плату, но если свой постраничный словарь – это надежная опора, Вы в нем уверенно ориентируетесь, можете многое вспомнить даже на ассоциативном уровне, то, будучи составленным «наемным работником», он превращается в еще один непонятный атрибут непонятного задания. Представьте себе спортсмена, который нанимает других людей, чтобы они ходили за него на тренировки, а потом удивляется, что не может сдать желанный норматив. Домашняя работа над переводом – такая же тренировка, которая дает неоценимый и недостижимый никакими другими средствами навык работы с иноязычным текстом.

3. Создайте список **ключевых слов** (**Key Words**):

3.1. просмотрите свой постраничный словарь и выберите из него 100 слов или выражений, которые можно отнести к **терминам**, **связанным с Вашей специальностью**, или к **научной лексике**;

3.2. в качестве русского эквивалента постарайтесь взять то **значение, в котором это слово употреблено в тексте**;

3.3. поставьте это слово или словосочетание, а также его русский эквивалент в начальную форму;

Примечание

Для имени существительного начальной формой является единственное число, именительный падеж; для русского прилагательного – полная форма, положительная степень, единственное число, мужской род, именительный падеж; для английского – положительная степень; для глагола – инфинитив.

3.4. оформите словарь в таблицу в соответствии с Приложением 5 на отдельной странице документа Word, в котором набирали перевод: в первой колонке таблицы находятся английские слова и выражения, во второй – транскрипция, в третьей – их русские эквиваленты;

3.5. транскрипцию можно как набирать на компьютере, используя дополнительные символы, так и распечатав список, написать её от руки;

Примечание

Дополнительные символы не всегда правильно отображаются при печати. Поэтому, прежде чем сдать отчет преподавателю, проверьте свой список ключевых слов, и, при необходимости, внесите исправления при помощи корректора.

3.6. все ключевые слова необходимо **выучить наизусть** и отчитаться преподавателю;

3.7. учить и сдавать слова можно порциями по 10-20 единиц, что настоятельно рекомендуется, или сразу всем списком (в любом случае желательно выполнить эту часть задания, не дожидаясь итогового контроля).

Примечание

Начать составление списка ключевых слов можно уже в ходе работы над постраничным словарем, корректируя его по мере пополнения и постепенно заучивая, - таким образом, слова будут запоминаться в контексте с использованием ассоциативных механизмов памяти.

4. Выполните **грамматическое задание** (**Grammar Task**):

4.1. согласовав с преподавателем тему своего грамматического задания (например, Past Perfect, Degrees of Comparison, Modal Verbs и т.п.), убедитесь, что владеете необходимым языковым материалом (т.е. знаете модель образования данной конструкции и понимаете, каким образом и в каких случаях она используется);

4.2. просмотрите текст, который Вы переводили и найдите не менее десяти употреблений данного грамматического явления;

4.3. на отдельной странице создайте таблицу из двух колонок, в первую скопируйте предложения, содержащие эту конструкцию, а во вторую – русский эквивалент этих предложений из Вашего задания;

4.5. в английских предложениях выделите нужную конструкцию подчеркиванием (Приложение 4).

5. Составление **пересказа переведенного текста (Summary)** – итоговый и очень важный пункт Вашей работы над отчетом. Именно это задание демонстрирует, насколько полно и правильно Вы разобрались в содержании переводимого фрагмента, и может ли английский текст, при необходимости, послужить Вам в качестве источника профессиональной информации. Для написания хорошего пересказа Вам потребуются навыки реферирования, полученные при работе с русскими текстами, ведь общие принципы одинаковы: необходимо определить тему, основные положения, ключевые определения и т.п. Кроме того, Вам придется согласиться с выводами автора или попытаться их оспорить.

Все эти требования выглядят устрашающе, однако Вы значительно упростите себе задачу, если будете следовать схеме пересказа, предлагаемой

в данном пособии. Эти клише Вы можете использовать при работе с текстами любого объема, в частности, во время экзамена или зачета по иностранному языку. Если Вы заучите их наизусть, то очень быстро это задание перестанет казаться невыполнимым, ведь буквально в 8-12 предложениях можно пересказать суть любого текста (конечно, при условии что Вы его поняли и смогли выделить главное).

Следует также помнить, что если Вам пришлось прочитать несколько статей, то пересказы надо составлять по каждой отдельно. То же касается работы над несколькими главами книги. При пересказе целостного фрагмента из книги сложность будет состоять в том, что переведенный Вами фрагмент может являть собой композиционно незаконченный отрывок. В этом случае Вы, вероятно, мало что сможете сказать об авторских выводах, что может повлиять на композицию пересказа. Можно сразу упомянуть об этом обстоятельстве в первых предложениях, а можно, при условии согласия Вашего ведущего преподавателя, в первом семестре начать работу над пересказом, а во втором, продолжив читать ту же книгу, завершить ее. Конечно, итоговый объем пересказа при таком подходе увеличится почти в два раза.

5.1**.** Начать необходимо с **названия текста и имени автора**:

5.1.1. если это отдельная статья, то

- The title of the article I am dealing with / under consideration / I am going to comment on is <*название*> by <*имя автора*>;
- The article I am dealing with / under consideration / I am going to comment on is entitled / headlined <*название*> by <*имя автора*>;

5.1.2. если речь идет о фрагменте книги, то

- The passage I am dealing with / under consideration / I am going to comment on is an extract from the book / work / study <*название*> by <*имя автора*>;
- The fragment <*название*> is taken from <*название*> by <*имя автора*>;

•This extract presents a chapter (two chapters and a half) from <*название*> by <*имя автора*>;

5.2. затем следует сообщить о **теме** текста:

•The author uses such key words as <*3-5 ключевых слов*> which make me think about the subject matter of the text / article / book / work / study. It is …;

•Key words like <*3-5 ключевых слов*> present the subject matter of the text / article / book / work / study which is …;

•The subject-matter of the text / article / book / work / study is …;

•The main idea of the text / article / book / work / study is …;

•The text / article / book / work / study is devoted to …;

•The text / article / book / work / study gives information about …;

•The text / article / book / work / study touches upon the problem concerning …;

•The text / article / book / work / study deals with …;

5.3. укажите на **стиль** текста и основные поднимаемые в нем **проблемы**:

•As for the style of the text / article / book / work / study it is thought / considered to be popular / scientific;

•The text / article / book / work / study sums up many burning problems of …;

•The text / article / book / work / study sums up such problems as …;

•The text / article / book / work / study contains no problems, it's just a description of …;

5.4. переходите к изложению **содержания**:

5.4.1. с чего начинает автор:

•At the beginning of the text the author dwells on / explains / mentions / points out / touches upon / introduces / comments on / reports about …;

•The text / article / book / work / study begins with mentioning / a comment on / the description of …;

5.4.2. основная часть:

•Then / after that / further on / next the author passes to / goes on to explain (comment on);

5.4.3. возможные вводные конструкции:

•To begin with …;

•Also …;

•In addition …;

•Besides …;

•Moreover …;

Примечание.

При изложении содержания можно использовать конструкцию пассивного залога, например, вместо The author shows that – It is shown that, The author mentions – It is mentioned that … и т.п.

5.4.4. окончание текста (если это отдельная статья или глава):

•At the end of the article / the chapter the author sums up by saying …;

•In conclusion the author …;

•The article / the chapter ends with …;

•Finally the author …;

5.5. выразите **свое мнение** по поводу текста:

•(To sum up / In conclusion I would like to say that) I find the text interesting / important / informative / useful / useless / of no value / boring / too difficult to understand because …;

•In my view / opinion …;

•I think / believe / guess / am sure / am convinced that …;

•I doubt if /that …;

•I suspect that … .

Примечание

Из предлагаемых вариантов клише следует выбрать для каждого пункта пересказа те, которые кажутся Вам наиболее подходящими. Можно составить на их основе свой индивидуальный шаблон и всегда

следовать ему, доведя навык до автоматизма. Можно применять каждый раз новые выражения, подходя к этому заданию творчески. В любом случае важно следить за грамматической согласованностью предложений. Пересказ можно сделать более развернутым и содержательным, обогатив его фрагментами из текста.

Пересказ, составленный просто из предложений, «вырванных» из исходного текста, без учета данных или аналогичных рекомендаций, не будет принят преподавателем. Помните, преподавателю Важно, чтобы Вы умели пользоваться иностранным языком, самостоятельно составлять предложения, а не просто заучивать отрывки текста наизусть.

V. Как происходит контроль домашнего чтения

В разделе II уже говорилось о том, что существует два вида контроля: промежуточный и итоговый. Обычно преподаватель заранее информирует студентов о сроках и требованиях. Желательно не игнорировать эти сроки и предъявлять итоги своей работы на каждом этапе вовремя.

1. Промежуточный контроль:

1.1. когда Вы придёте отчитываться первый раз, не забудьте взять с собой оригинал текста или его ксерокопию и постраничный словарь, насколько он готов к этому времени, т.к. в противном случае Вашу работу не примут;

1.2. желательно, чтобы Вы уже начали составлять список ключевых слов и были готовы сдать наизусть 10-20 терминов;

1.3. преподаватель проверит, как Вы выучили слова, поинтересуется, сколько страниц Вы перевели, и предложит прочитать и перевести вслух несколько фрагментов из их числа (Вы можете пользоваться постраничным словарем, если возникают затруднения);

Примечание

Если Ваш постраничный словарь достаточно полный (т.е. выписаны все слова и выражения, вызывающие затруднения), хорошо структурирован (т.е. четко отмечены номера страниц и абзацев), и Вы в нем хорошо ориентируетесь (т.е. выполняли перевод самостоятельно), то он станет для Вас очень удобным инструментом и облегчит работу.

1.4. Вам предложат продемонстрировать готовый переведенный фрагмент (распечатанный параллельный перевод), который будет проверен преподавателем;

1.5. преподаватель отметит в своем журнале объем сданного Вами материала, даст рекомендации по дальнейшей работе и назовет срок следующего промежуточного или итогового контроля.

2. Итоговый контроль:

2.1. Вам предложат отчитаться по оставшемуся материалу по тому же принципу, что и в п. 1;

2.2. Вы сдадите на проверку полностью готовый письменный отчет (требования к оформлению подробно описаны в разделе II);

2.3. в случае неудовлетворительного результата Вам предложат внести в отчет коррективы к определенному сроку.

VI. Что нельзя делать при работе над домашним чтением.

Подведем итоги и вспомним наиболее характерные ошибки при работе над домашним чтением.

1. Нельзя искать значение слова в словаре, не определив, какая это часть речи в данном тексте. Давайте восстановим в памяти название русских частей речи и их обозначения в английском языке.

существительное	noun (n)
глагол	verb (v)
прилагательное	adjective (adj)
наречие	adverb (adv)
союз	conjunction (conj)
предлог	preposition (prep)
местоимение	pronoun (pron)

2. Нельзя смотреть слова в словаре «списком», без контекста. Выписывание всех незнакомых слов из нескольких страниц текста, а затем списывание подряд их значений из словаря не принесет положительного результата, так как для английского языка характерна многозначность слов, схожие словоформы для разных частей речи, а также наличие большого количества фразеологических словосочетаний, в которых искомое слово может иметь кардинально противоположный смысл.

3. Нельзя выписывать первое значение слова из словаря. Необходимо просмотреть все его значения и определить, какое из них лучше всего соответствует контексту.

4. Не стоит пользоваться миниатюрными словарями, так как в них содержится ограниченное количество слов и их значений.

5. Нельзя игнорировать правила словообразования. Если Вы знакомы с этими правилами, то, посмотрев одно слово в словаре, вы сможете узнать и перевести множество однокоренных слов. Например, use (v), us**er** (n), use**ful** / use**less** (adj), use**ful***ly* / use**less***ly* (adv), etc.

6. Нельзя просто перевести текст с помощью электронного переводчика и сдать его преподавателю на проверку. Необходимо помнить, что системы машинного перевода не являются совершенными системами. Зачастую текст, переведенный с их помощью, содержит большое количество лексических и грамматических ошибок, а предложения просто не имеют смысла. Преподаватель сразу же определит, что работа была выполнена не самостоятельно, и вправе вернуть такой перевод без проверки с выставлением неудовлетворительной оценки.

В качестве примера авторы хотят привести примеры из работ студентов, пытавшихся сдать отчет по домашнему чтению, выполненный при помощи электронного переводчика.

There has been, and is, a reverence for the ground or soil.	Там был, и есть, почтение к земле или почвы.
Roots anchored in soil enable growing plants to remain upright.	Корни якорь в почву включить выращивания растений, чтобы оставаться в вертикальном положении.
This student’s mother teaches English at the university.	Этот студент мать преподавателей английского в университете.
What is this job like?	Что это за работа, как?
Employers are especially looking for chemists and material scientists who have a master’s or PhD degree.	Работодатели особенно глядя на химики и материалы ученых, имеющих степень магистра или доктора наук степень.
Finally, during stage four, there is the production of the ear or the grain.	Наконец, во время четвертой стадии, то производство колоска или зерна.

7. Нельзя найти две одинаковые книги на английском и русском языке и, распечатав оба варианта, выдать перевод профессионала за

свой труд. Необходимо помнить, что профессиональный перевод текста не всегда полностью соответствует оригиналу. Переводчик, работая над книгой, может изменять структуру предложений, изменять языковые реалии, использовать фразеологизмы и устойчивые словосочетания, характерные для языка перевода, а не языка оригинала. Кроме того, художественный или научно-популярный стиль, которым будет написана книга, несколько отличается от того стиля, которым студенты обычно переводят иностранные тексты. Более того, преподаватель также имеет доступ к профессиональным переводам книг, а полное совпадение перевода, вплоть до всех знаков препинания, включая двоеточия и тире, выполненное двумя разными людьми, просто невозможно. В случае плагиата преподаватель также вправе считать работу неудовлетворительной.

В качестве примера хочется привести небольшой отрывок из книги Артура Конан Дойля «Голубой карбункул». *Курсивом* отмечены места, где перевод не совпадает с оригиналом.

The Adventure of the Blue Carbuncle	Голубой карбункул
I had *called upon* my *friend* Sherlock Holmes upon the *second morning after Christmas*, with the intention of *wishing him the compliments of the season*. He was lounging upon the sofa in a *purple* dressing-gown, a pipe-rack within his reach upon the right, and a pile of crumpled morning papers, evidently newly studied, *near at hand*. Beside the couch was a *wooden* chair, and on the angle of the back hung a very seedy and disreputable hard-felt hat, *much the*	*На третий день Рождества зашел* я к Шерлоку Холмсу, чтобы *поздравить его с праздником*. Он лежал на кушетке в *красном* халате; по правую руку от него была подставка для трубок, а по левую – груда помятых утренних газет которые он, видимо, только что просматривал. Рядом с кушеткой стоял стул, на его спинке висела сильно поношенная, потерявшая вид фетровая шляпа. *Холмс, должно быть очень внимательно изучал эту*

worse for wear, and cracked in several places. A lens and a forceps lying upon the seat of the chair suggested that *the hat had been suspended in this manner for the purpose of examination*.	*шляпу*, так как тут же на сиденье стула лежали пинцет и лупа.
"*You are engaged*," said I; "perhaps I interrupt you."	– *Вы заняты?* – сказал я. – Я вам не помешал?
"Not at all. I am glad to have a friend with whom I can discuss my results. The *matter* is a perfectly trivial one" – he jerked his thumb in the direction of the *old* hat – "but there are points in connection with it which are *not entirely devoid of interest and even of instruction*."	– Нисколько, – *ответил он.* – Я рад, что у меня есть друг, с которым я могу обсудить результаты *некоторых моих изысканий*. *Дельце* весьма заурядное, но с этой вещью, – он ткнул большим пальцем в сторону шляпы, – связаны кое-какие *любопытные и даже поучительные события*.

Авторы данного пособия надеются, что оно поможет студентам в выполнении заданий по домашнему чтению, ответит на большинство возникающих у них вопросов, а также, может быть, привьет студентам любовь к чтению иноязычной литературы. Авторы будут благодарны всем, кто, прочитав это пособие аудитория) и оставит и не найдя ответов на свои вопросы, придет на кафедру английского языка для профессиональной деятельности (3 корпус, ауд. 212, email: english-chair@mail.ru) свои пожелания или замечания.

Приложение 1. Образец титульного листа

Министерство образования и науки РФ

Федеральное государственное бюджетное образовательное учреждение высшего профессионального образования

“Ульяновский государственный университет”

Институт международных отношений

Факультет лингвистики, межкультурных связей и профессиональной коммуникации

Кафедра английского языка для профессиональной деятельности

Перевод

(Название книги на английском языке, автор, год издания, номера переведенных страниц / или www. откуда скачали книгу)

Выполнил:

Студент № группы

ФИО

Проверил:

Должность преподавателя

(доцент / ст. преподаватель / ассистент каф. АЯПД

ФИО

Дата сдачи:

Оценка:

Ульяновск, 2014

Приложение 2. Образец параллельного перевода. *История*

Diplomacy after the civil war.	Дипломатия после гражданской войны
The overriding concerns of the time were the development of industry, settlement of the West, and domestic politics. Compared to these, foreign relations simply were not important to the vast majority of Americans.	Первостепенными задачами того времени были: развитие промышленности, обустройство Запада страны и внутренняя политика. По сравнению с этими проблемами, международные отношения просто не являлись значимыми для подавляющего большинства американцев.
The major issues of foreign affairs stemming from the Civil War were settled within a few years of its end, after which few conflicts or major issues arose to trouble the American people. A mood of isolation settled upon the United States, favored since the War of 1812 with what one historian called «free security»: wide oceans as buffers on either side, the British navy situated between America and the powers of Europe.	Основные вопросы внешней политики, возникшие как следствие Гражданской войны, были урегулированы в течение нескольких лет после ее окончания, после чего, американский народ сталкивался лишь с единичными конфликтами или проблемами. В США установилось настроение изоляционизма, чему, со времен войны 1812 года, благоволила (по словам одного историка) «бесплатная безопасность»: огромные океаны как буферы со всех сторон; флот Англии, располагавшийся между Америкой и европейскими державами.

Приложение 2. Образец параллельного перевода. *Экология*

Comparing species	Сравнивая виды.
How are the differences between humans and other organisms reflected in our genomes? How similar are the numbers and types of proteins in humans, fruitflies, worms, plants and yeast? And what does all of this tell us about what makes a species unique?	Как различия между человеком и другими организмами отражены в наших геномах? Насколько сходны число и виды белков человека, дрозофил, червей, растений и дрожжей? И что все это говорит нам о том, что делает виды уникальными?
An obvious place to start our comparison is the total number of genes in each species. Here is a real surprise: the human genome probably contains between 25,000 and 40,000 genes, only about twice the number needed to make a fruitfly, worm or plant. We know that there is a higher degree of “alternative splicing” in humans than in other species. In other words, there are often many more ways in which a gene’s protein-coding sections (exons) can be joined together to create a functional messenger RNA molecule, ready to be translated into protein.	Очевидно, что сравнение нужно начать с общего числа генов у каждого вида. Здесь нас ожидает настоящий сюрприз: геном человека вероятно содержит от 25000 до 40000 генов, всего лишь в 2 раза больше, чем у дрозофилы, червя или растения. Мы знаем, что существует высокая степень «альтернативного сплайсинга» у человека по сравнению с другими видами. Другими словами, часто существует намного больше способов соединения, кодирующих белок участков генов (экзонов) друг с другом, чтобы создать функциональную молекулу матричной РНК, готовой для трансляции в белок.

Приложение 2. Образец параллельного перевода. *Журналистика*

The Corporate “Free Press”	Корпоративная “Свободная пресса”
Even the word “media” is problematic. It is the plural of the word medium, which can be denned as “the intervening substance through which impressions are conveyed to the senses”.	Даже само слово «медиа» неоднозначно. Это множественное число от слова «медиум», которое можно определить как «дополнительную субстанцию, через которую впечатления преобразуются в чувства».
News organizations would have us believe that they transmit information in a similarly neutral, natural way. They represent themselves as self-evidently dispassionate windows on the world. Thus, while there is plenty of discussion about what appears in these windows, there is next to no discussion about who built them, about what their goals and values might be.	Службы новостей хотели бы заставить нас поверить в то, что они передают информацию нейтральным, естественным образом. Они представляют себя как самоочевидные, беспристрастные окна в окружающий мир. Таким образом, в то время как существуют множество дискуссий о том, что появляется в этих окнах, почти отсутствуют дискуссии о том, кто создает их, о том, каковы могут быть их цели и ценности.
And yet consider two facts: 1) contemporary world is dominated by giant, multinational corporations; 2) the media system is itself made up of giant corporations.	А теперь рассмотрим два факта: 1) в современном мире доминируют гигантские транснациональные корпорации; 2) система СМИ, освещающая сама создана из гигантских корпораций.

Приложение 2. Образец параллельного перевода. *Юриспруденция*

Young Killers	Юные убийцы
One of the most disturbing things about the United States today is the violent behavior of our youth. More and more young people are arrested for homicide, aggravated assault and the use of weapons, and the age of which they commit considerable numbers of such crimes continues to move downward.	Один из наиболее волнующих фактов в США сегодня – это агрессивное поведение молодежи. Всё больше и больше молодых людей подвергаются аресту за убийство, нападение с применением физического насилия и применение оружия, и возраст, в котором они совершают огромное количество подобных преступлений, продолжает снижаться.
Homicide among black males is such that we may expect about one in twenty to be a victim of homicide. And since like tends to kill like we may expect a large share of youthful victims to be killed by their peers.	Процент убийств среди темнокожих мужчин таков, что предположительно каждый двадцатый из них может стать жертвой убийства. И так как тенденция к убийствам растет, предполагается, что большая часть молодых людей становятся жертвами своих сверстников.
Fortunately females are less violent. Since women are less likely to kill, they are also less likely to be killed.	К счастью женщины менее склонны к насилию. Так как женщины с гораздо меньшей вероятностью склонны убивать, они с меньшей же вероятностью, становятся жертвами убийств.

Приложение 2. Образец параллельного перевода. *Психология*

What is Child Development?	Что означает развитие ребенка?
Child development involves the scientific study of changes in the child's biological, social, cognitive, and emotional behavior across the span of childhood. Two central questions about development concern all child psychologists.	Развитие ребенка включает в себя научное исследование изменений в биологическом, социальном, познавательном и эмоциональных аспектах поведения в детстве. Два основных вопроса о развитии детей интересуют всех детских психологов.
First, how do children change as they develop? Second, what are the determinants of these developmental changes? Child psychology involves both the description and explanation of changes in children's development. It is not enough simply to be able to state that older children can learn to solve logic problems better than younger children, to detail how an infant's ability to grasp objects improves, or to describe children's increasing skill in understanding the feelings of others.	Во-первых, как дети изменяются в процессе развития? Во-вторых, каковы признаки этих изменений в развитии? Детские психологи описывают и объясняют изменения в развитии детей. Недостаточно просто установить тот факт, что взрослые дети могут решать логические задачи лучше, чем дети, которые младше их. Недостаточно только установить, как развивается способность маленького ребенка брать предметы или описать развивающуюся способность людей понимать то, что чувствуют другие.

Приложение 3. Образец выполнения пересказа.

Summary

The text I'm going to comment on is a fragment from the book Deborah Swiss and Judith Walker "Women and the Work". **I know little about the authors but I would suppose that they are interested in** psychology because they try to solve some psychological problems which women face in their lives.

So the subject matter of the book is to describe women's problems and to find workable solutions to the balancing act between career, marriage and children.

The information sums up many burning problems concerning career and children, part-time careers, full-time parents and so on. **On the one hand** this book is mainly devoted to mothers. **But on the other hand** it will be also important for husbands and employers.

So let me tell you some words about the plot of the information. First of all the author underlines that all the material described in the book is based on the analysis of 902 surveys and 52 personal interviews. **Then the authors point out** that nowadays many women have to choose between a profession and a family.

The authors mention that there exists such a paradox: the term working father is not even in our vocabulary, because society usually does not demand fathers to be both good parents and professionals. As for the term working mother it seems clear to everyone.

The next chapters are devoted to separate women and their life choices. **In chapter one, we learn about** Laura Tosi and her ability to combine career and family successfully. Today she is one of a few female orthopedic surgeons in the country and a mother of two children. But not all examples are so optimistic. Sometimes women are penalized for their choice to have children. It is a myth that it is possible to have everything and at once. Sometimes it is necessary to sacrifice something otherwise you will get nothing.

In conclusion the authors give some advice for women who want to coincide their careers and families.

In my conclusion it is necessary to say that the problems discussed in the book are very up-to-date. **I think that** in our country women have the same problems with families and careers as in the United States. I think that this book will be interesting first of all for women. While reading they will be able to learn how to make individual decisions, how not to become angry and frustrated while coming along different problems. **This book may also help** them understand that they are not alone, that it is possible to go public with their personal dilemma. I believe that psychologist would also find it interesting because some strategies proposed in the book can be rather useful in their work. **What is more** this book may also be helpful for student who study psychology as it can teach them how to conduct a survey and what questions to include in it.

As for me I am glad that I read at least one chapter of this book, because I think it may help me in my future work.

Приложение 4. Образец выполнения грамматического задания.

Modal Verbs

1. You must develop your language skills if you want to become a god interpreter.	Если ты хочешь стать хорошим переводчиком, ты должен развивать свои языковые умения.
2. The dean is to cope with a great deal of administrative work.	Декан должен справляться с большим количеством административной работы.
3. You should hire high qualified personal if you want your company to develop.	Если ты хочешь, чтобы твоя компания развивалась, тебе следует нанять хорошо квалифицированный персонал.
4. You cannot have solved this puzzle yourself.	Не может быть, чтобы ты решил эту задачу самостоятельно.
5. He would regret he had not accepted their proposal.	Он, бывало, сожалел, что не принял их предложение.
6. You ought to include this topic in your report.	Тебе следует включить эту тему в свой доклад.
7. You needn't make a secret of this information, it's available to everyone.	Тебе не нужно делать секрета из этой информации, она доступна всем.
8. He may not have graduated from the University, he is so stupid.	Он не мог окончить университет, он такой глупый.
9. We'll have to increase the number of problems under consideration.	Нам придется увеличить число рассматриваемых проблем.
10. How dare you appear here again?	Как ты посмел здесь снова появиться?

Приложение 4. Образец выполнения грамматического задания.

Condition

1. If he had not supported us we would have failed.	Если бы он нас не поддержал, мы бы провали дело.
2. If I were you I would demand salary increase.	На вашем месте, я бы требовал повышения зарплаты.
3. If he had really suffered he wouldn't have spent so much money.	Если бы он действительно страдал, он бы не тратил столько денег.
4. If my expectations come true, we will be able to go to Spain this summer.	Если мои ожидания оправдаются, мы сможем поехать в Италию этим летом.
5. If there were a breadwinner in their family the children wouldn't have to work.	Если бы в их семье был кормилец, детям не пришлось бы работать.
6. Had there not been serious obstacles he would have fulfilled the task.	Не будь серьезных преград, он бы выполнил задачу.
7. If he had more free time his research would evolve into a book.	Если бы у него было больше свободного времени, его исследование переросло бы в книгу.
8. If the employer had paid more money the workers would not have gone on a strike.	Если бы работодатель платил больше денег, рабочие не вышли бы на забастовку.
9. Were there no gap between generations parents and children would understand each other better.	Не будь разрыва между поколениями, родители и дети понимали бы друг друга лучше.
10. If her children didn't help her she would not manage the household.	Если бы дети не помогали ей, она бы не смогла управлять домашним хозяйством.

Приложение 4. Образец выполнения грамматического задания.

Infinitive and Infinitive Constructions

1. I consider success to be one of the main components of every enterprise. (Complex Object)	Я считаю успех одним из главных компонентов любого предприятия.
2. He told us an anecdote in order to reduce tension.	Чтобы снизить напряжение, он рассказал нам анекдот.
3. Parents were proud of their son as he was considered to be the best pupil in class. (Complex Subject)	Родители гордились своим сыном, так как он считался лучшим учеником в классе.
4. His had a very strange ability always to win.	У него была очень странная способность всегда выигрывать.
5. Mother made her children come home by 8 p.m. as she believed these restrictions to be very useful. (Complex Object)	Мать заставляла своих детей приходить домой до 8 часов вечера, так как она считала эти ограничения очень полезными.
6. To participate in the conference you must pay the first fee.	Чтобы участвовать в конференции, вы должны заплатить первый взнос.
7. To become a good supervisor you should be experienced enough.	Чтобы стать хорошим руководителем, вы должны быть достаточно опытным.
8. Everyone believed him to be fired without any cause. (Complex Object)	Все полагали, что его уволили без какой-либо причины.
9. He decided to take part in competitions to get everyone's attention.	Он решил принять участие в соревнованиях, чтобы привлечь всеобщее внимание.
10. TV, newspapers, magazines and radio are known to be examples of mass media. (Complex Subject)	Известно, что телевидение, газеты, журналы и радио являются примерами средств массовой информации.

Приложение 4. Образец выполнения грамматического задания.

Participle I, Participle II

1. The last issue of this magazine devoted to the wedding was very popular. (Participle II)	Последний номер этого журнала, посвященный свадьбе, пользовался большой популярностью.
2. One thousand people responded to our survey including men and women. (Participle I)	Тысяча людей, включая мужчин и женщин, приняли участие в нашем исследовании.
3. Stepchildren adopted many years ago didn't even know where they had been born. (Participle II)	Приемные дети, которых усыновили много лет тому назад, даже не знали, где они родились.
4. Tom knew that he would be punished for the broken vase. (Participle II)	Том знал, что его накажут за разбитую вазу.
5. Working women face many complicated tasks. (Participle I)	Работающие женщины сталкиваются со множеством сложных проблем.
6. Being a real genius his perception differs much from our perception. (Participle I)	Так как он является настоящим гением, его восприятие намного отличается от нашего.
7. Have you ever seen that extraordinary man standing near the window? (Participle I)	Ты когда-нибудь видел того необычного человека, стоящего около окна.
8. He felt in a fury trying to solve the equation. (Participle I)	Он впал в ярость, пытаясь решить уравнение.
9. Having launched his own business he was very happy to gain the first profit. (Participle I)	Начав свой собственный бизнес, он был очень рад получить первую прибыль.
10. The lawyer spoke to his client not looking at him. (Participle I)	Юрист разговаривал со своим клиентом, не глядя на него.

Приложение 4. Образец выполнения грамматического задания.

Gerund

1. What is the reason for your being late again?	Какова причина того, что вы снова опоздали.
2. The manual can give you some strategies for attaining a peaceful alliance between parents and children.	Это пособие может предложить вам некоторые стратегии достижения мирного союза между родителями и детьми.
3. You should be more attentive in managing your obligations.	Тебе следует быть более внимательным в выполнении своих обязанностей.
4. He found consolation in realizing that he was not the only person who had failed to pass the exam.	Он нашел утешение, осознав то, что он был не единственным человеком, который не сдал экзамен.
5. Mary was very happy that her father had agreed to her being married.	Мэри была очень рада, что ее отец согласился на то, чтобы она вышла замуж.
6. Excuse my hesitation while making this decision.	Извини за то, что я сомневался при принятии этого решения.
7. This claim is worth while satisfying.	Эту претензию стоит удовлетворить.
8. I'm against of our rights being violated.	Я против того, чтобы нарушались наши права.
9. He couldn't stand devastating the area where he had spent his childhood.	Он не мог вынести разорения района, где он провел свое детство.
10. You can always rely on the older generation giving you a piece of advice.	Ты всегда можешь рассчитывать на то, что старшее поколение даст тебе совет.

Приложение 5. Образец выполнения списка ключевых слов.

Key Words

№	Английское слово	Транскрипция	Перевод
1	to develop	[di'veləp]	развивать
2	available	[ə'veiləbl]	доступный
3	a breadwinner	['bredwinə]	кормилец
4	expect	[ik'spekt]	ожидать
5	an employer	[im'ploiə]	работодатель
6	success	[sək'ses]	успех
7	to support	[sə'po:t]	поддерживать
8	a survey	['sə:vei]	исследование
9	a dean	[di:n]	декан
10	a supervisor	['sju:pəvaisə]	руководитель
11	a cause	[ko:z]	причина
12	attentive	[ə'tentiv]	внимательный
13	a lawyer	['lo:jə]	юрист
14	to avoid	[ə'void]	избегать
15	a reason	['ri:zn]	причина
16	applied fields		прикладные области
17	mental processes		умственные процессы
18	behaviour	[bi'heivjə]	поведение
19	related to		связана с
20	in nature		по своему характеру

Приложение 6. Список сайтов.

1.

Сайт кафедры английского языка для профессиональной деятельности

http://www.ulsu.ru/com/faculties/elhs/

2.

Социальная работа

1. http://www.naswdc.org/ - официальный сайт национальной ассоциации социальных работников. Много статей по различным актуальным темам и проблемам.
2. http://www.bls.gov/ooh/Community-and-Social-Service/Social-workers.htm - Сайт министерства по труду США. Хорошие тексты общего характера по обязанностям, трудоустройству, зарплате соцработников и т.д.
3. http://www.socialworker.com/home/index.php- сайт для студентов и социальных работников. Много полезной современной информации.
4. http://www.basw.co.uk/social-work-careers/ - сайт британской ассоциации соцработников.

Психология

1. http://www.psychologytoday.com/ - огромный выбор текстов по различным отраслям психологии
2. http://www.psychology.org/ - Энциклопедия психологии.

История

1. http://www.rulit.net/tag/history/en/1/date Книги в удобном для скачивания формате.
2. http://www.history.com/ - Посвящен как персоналиям, так и историческим событиям. Есть видеоматериалы.
3. http://www.bbc.co.uk/history/0/ - сайт BBC по истории с A-Z указателем. Затронуто огромное количество разнообразных тем. Можно найти текст на любой вкус.

4. http://www.besthistorysites.net/ - большое количество ссылок на книги по истории.

Социология

1. http://www.sociology.org/ - Социологический журнал

Юриспруденция

1. http://www.law.northwestern.edu/jclc/ - журнал по уголовному праву и криминалистике. Статьи в pdf.

Экология

1. http://www.eoearth.org/article/Ecology - энциклопедия по экологии. Большое количество статей на разнообразные экологические темы.
2. http://www.journalofecology.org/view/0/index.html - журнал по экологии, рассматриваются современные экологические проблемы.
3. http://www.globalissues.org/issue/168/environmental-issues - глобальные экологические проблемы.
4. http://www.nrdc.org/issues/ - сайт Совета по защите природных ресурсов.

Приложение 7. Тексты для чтения

ЮРИДИЧЕСКИЙ ФАКУЛЬТЕТ

Crime and Punishment

Introduction

Within a broad spectrum of cultural and historical variations, crime constitutes the intentional commission of an act usually deemed socially harmful or dangerous and specifically defined, prohibited, and punishable under the criminal law. Most countries have enacted a criminal code in which all of the criminal law can be found, although English law--the source of many other criminal law systems--remains uncodified. The definitions of particular crimes contained in a code must be interpreted in the light of many principles, some of which are not expressed in the code itself. The most important of these are related to the mental state of the accused person at the time of the act that is alleged to constitute a crime. Crimes are classified by most legal systems for purposes such as determining which court has authority to deal with the case. Social changes often result in the adoption of new criminal laws and the obsolescence of older ones.

The purpose of punishing offenders has been debated for centuries. A variety of often conflicting theories are held, and in practice each is followed to some extent. Prison is not the most common penalty for crime--a wide variety of punishments that do not involve incarceration have developed, including financial sanctions, such as fines, and schemes for service to the community in general or to the victim in particular. Juveniles are usually dealt with by courts set aside exclusively for the prosecution of young offenders.

The prison systems of most countries are subject to many problems, especially overcrowding, but the recognition by some legal systems that prisoners have rights that the courts can enforce has led to some improvements. The death penalty is now rare in Western countries, although it has been reinstated in some parts of the United States after a period of disuse.

The concept of crime

CRIMINAL CODES AND OTHER LEGAL FORMULATIONS

Crime is whatever conduct the laws of a particular jurisdiction designate as criminal, and there are many differences from one country to another as to what behaviour is prohibited. Conduct that is lawful in one country may be criminal in another, and activity that amounts to a trivial infraction in one country may constitute a serious crime elsewhere. Changing times and social attitudes may lead to changes in the criminal law, so that behaviour that was once criminal becomes lawful. Abortion, once prohibited except in the most unusual circumstances, has become lawful in many countries, as has homosexual behaviour in private between consenting adults, which was once a serious offense. Suicide and attempted suicide, once criminal, have also been removed from the scope of the criminal law in many countries. Nonetheless, the trend generally is to increase the scope of the criminal law rather than to reduce it. It is more common to find that statutes create new criminal offenses than that they abolish old ones. New technologies give rise to new opportunities for their abuse, which in turn give rise to legal restrictions; just as the invention of the motor vehicle led to the development of a whole body of criminal laws designed to regulate its use, so the widening use of computers has created the need to legislate against a variety of new abuses and frauds--or old frauds committed in new ways.

The English-speaking world.

In most countries the criminal law is contained in a single statute, known as the criminal code or penal code. Although the criminal codes of most English-speaking countries are derived from English criminal law (in many cases with substantial modifications), England itself has never had a criminal code. English criminal law still consists of a collection of statutes of varying age (the oldest still in force being the Treason Act, 1351) and a set of general principles that are chiefly expressed in the decisions of the courts (case law). The absence of a criminal code in England is not for want of effort; since the early 19th century there has been a series of attempts to reduce the English criminal law to a code. The first effort, which occurred between 1833 and 1853, was by two panels of criminal law commissioners, who systematically surveyed the prevailing state of

the criminal law. They were confronted by a vast number of often overlapping and inconsistent statutes. Determining precisely what the law provided on any particular topic was enormously difficult; the existence of different statutes covering the same conduct, often with widely varying penalty provisions, permitted wide judicial discretion and inconsistency of punishment. The English criminal law commissioners drew up a number of draft codes that were presented to Parliament but not enacted. Eventually, the resistance of the judiciary led to the abandonment of the movement toward codification, and instead there was a consolidation of most of the statutory criminal law in 1861 into a number of statutes--the Larceny Act, 1861, the Malicious Damage Act, 1861, and the Offences Against the Person Act, 1861, were the most important of these. As these statutes were consolidations rather than codifications, they preserved many of the difficulties of the earlier legislation, which was in effect reproduced in the form of a single statute without substantial change. The Offences Against the Person Act is still largely in force; the others have been replaced by more modern provisions.

Interest in codification was not limited to England; a similar process ensued in India, then under British rule, and a criminal code in the true sense was written during the 1830s and eventually enacted in 1860. It remains substantially in force in India and Pakistan and in certain parts of Africa that were once British colonies. The effort to produce a criminal code in England was resumed in 1877, and a further Criminal Code Bill was presented to Parliament in 1879-80. This draft code, largely the work of the celebrated legal author and judge James Fitzjames Stephen, was not enacted in England largely because Parliament was preoccupied with other matters at the time, but it was subsequently enacted in Canada as the Canadian criminal code in 1892 and in several Australian states and other British colonies.

Reform of the criminal law became one of the interests of the U.S. states in the period following the Revolution, and in the early 1820s a comprehensive draft code was prepared for Louisiana, although it was never enacted. Other states moved to codify their criminal laws, and New York enacted a criminal code in

1881, setting an example that was eventually followed by the majority of states. Because criminal law is primarily a matter for the individual states (in contrast to Canada, for example, where the national parliament enacts the criminal code for the country as a whole), there has been considerable variation in the content of the criminal code from one state to another. In 1950 the American Law Institute began more than a decade of effort that was to lead eventually to the publication of the Model Penal Code in 1962. This was an attempt to rationalize the criminal law in relation to modern society and to establish a logical framework for defining offenses and a consistent body of general principles on such matters as criminal intent and the liability of accomplices. The Model Penal Code had a profound influence on the revision of many individual state codes over the following 20 years; the code itself was never enacted completely, but it inspired and influenced a long period of criminal code reform.

Other systems.

Criminal offenses in the legislation of modern African countries are, with the exception of Sierra Leone and some southern African states, now defined in criminal or penal codes. As far as the common-law countries are concerned, this is a radical departure from the originally uncodified criminal law of England, on which these codes are largely based. Because of their origins, these codes reflect the penal assumptions of the original colonial metropole. The only concessions to local African values or problems are, first, the inclusion of legislation against various customary practices, notably witchcraft; second, the extension of the criminal law in states with a planned economy to cover economic crimes against the state; and, third, as a consequence of the soaring rate of some kinds of crime, special provision for offenses such as armed robbery. Special tribunals, not subject to the ordinary rules of procedure, have been established in many countries to deal with such offenses; similarly, special tribunals and commissions with punitive powers have been set up to investigate the assets or the misdeeds of former rulers displaced by coups d'état or revolutions.

The states of the world that have large Muslim communities fall into four main categories in matters of criminal law. First, there are those that have an English colonial past and that have in the main adopted English criminal law and procedure, such as Pakistan, Bangladesh, Jordan, and some of the Persian Gulf states. The second group comprises those states that came under French colonial influence and adopted French laws; these are the states of the Maghrib and North Africa, including Egypt, and also Syria and Iraq. The third group comprises those states that were relatively little influenced by a colonial presence and that retain Islamic law (called Shari'ah) with few or no reforms; these include Saudi Arabia and Iran (the last shah of Iran had reformed a large amount of the law, building on previous colonial laws, but this was almost totally eradicated following the revolution of 1979). The fourth group comprises those states, such as India and some East African countries, where Muslims are a minority community.

GENERAL PRINCIPLES OF CRIMINAL LAW

Rule against retroactivity.

Despite differences of form and detail, the general principles of criminal law have much in common throughout the English-speaking world. One widely accepted principle is the rule against retroactivity—an individual may not be punished for an action that was not designated a crime at the time it was carried out. This rule, which restricts the authority of the judges to declare new offenses (although not necessarily to expand the scope of old ones by interpretation) has not always been accepted in England; as late as 1960 the House of Lords in its judicial capacity claimed that the courts retained the authority to recognize new offenses as social needs changed and declared that it was criminal to publish a directory of prostitutes. In Scotland (the legal system and criminal courts of which are totally separate from those of England) the claim is still maintained. A shopkeeper who supplied equipment for glue sniffing to children was convicted and imprisoned, notwithstanding that the law contained no express provision prohibiting such conduct.

Determining which particular conduct constitutes a crime usually requires an examination of the terms of the relevant provisions of the code or statutory provisions (a few offenses in English law have not been defined in statute), but these must be interpreted in the light of a number of general principles. The most important of these is that an individual is not normally to be held guilty of a crime unless he intended or foresaw the consequences of his action or was aware of the circumstances that make it criminal. In advanced Western societies, legal codes frequently recognize mental abnormality such as schizophrenia, mental retardation, or paranoia as at least mitigating, if not absolving, factors, though the claim of insanity may be contested. Criminal law as a general rule does not punish accidental or negligent behaviour. This principle, known as the mens rea ("guilty mind") principle, is subject to many exceptions and qualifications. For many offenses (offenses of strict liability) it is abandoned completely, and in other cases it is allowed only a limited scope. If the offense is one that requires proof of intention or knowledge on the part of the accused, the court or jury must be satisfied that the accused himself had the necessary intention or knowledge at the time when he committed the act that constitutes the crime--it is not normally sufficient to prove that any ordinary person in his place would have realized what was likely to happen. This is a difficult matter to prove, and in practice the members of the jury can be guided only by what they would have intended themselves, if they can imagine themselves in the same situation as the accused, and by the accused person's explanations of his behaviour. The fact that an individual had been drinking before committing a crime is not in itself a defense, but it may in some cases be evidence that the accused person did not have the intention that the law requires for the offense with which he is charged. It is no defense for an accused person to say that because he had been drinking he acted out of character and did things that he would not have done if sober, but it may help him to persuade the court that he did not realize what the consequences of his actions would be if he shows that he was affected by drink. Provocation is not generally a defense to a criminal charge, except in the case of murder; in a murder

case evidence of a high degree of provocation (in English law, sufficient to provoke a reasonable person to act in the same way as the accused did) results in a verdict of manslaughter, even if the killing was intentional. One very rare condition that gives a complete exemption from criminal liability is a form of involuntary conduct known as automatism. This is a state (such as sleepwalking or certain effects of concussion) in which the conscious mind does not control the bodily movements, and thus the individual cannot be held accountable for potential consequences, however serious they may be.

Criminal responsibility.

Criminal responsibility is not limited only to those who perform the criminal acts themselves. As a general principle, anyone who "aids and abets" the perpetrator by encouraging or in any way knowingly helping him (for instance, by providing information, implements, or practical help) is an accomplice and is considered equally guilty. Those who actually perform the criminal act (e.g., wielding the weapon that strikes the fatal blow) are called principals in the first degree; those who assist at the time of the commission of the offense (e.g., holding the victim down while the principal in the first degree strikes the blow) are principals in the second degree; and those who assist before the crime takes place (e.g., by lending the weapon or by providing information) are accessories before the fact. As a general rule, all are equally responsible in the eyes of the law and liable to the same punishment. In many cases the accessory before the fact will be considered more culpable--if, for instance, he has instigated the offense and arranged for it to be committed by an associate. In some cases the person who actually performs the act that causes the crime is completely innocent of all intent--for instance, the nurse who administers to a patient, on the doctor's instructions, what she believes to be medicine but what is in fact poison. In this situation the person who carries out the act is an innocent agent and is not criminally responsible; the person who causes the innocent agent to act is the principal in the first degree. The accessory after the fact is one who helps a felon to evade arrest or conviction, by, for example, hiding him or destroying evidence. In England and some other

jurisdictions the expression is no longer in use, as specific offenses have been enacted to deal with this kind of behaviour.

Classification of crimes

GENERAL CONSIDERATIONS

Most legal systems find it necessary to divide crimes into categories for various purposes connected with the procedure of the courts--determining, for instance, which kind of court may deal with which kind of offense. The common law originally divided crimes into two categories--felonies (the graver crimes, generally punishable with death, which resulted in forfeiture of the perpetrator's land and goods to the crown) and misdemeanours (for which the common law provided fines or imprisonment). There were many differences in the procedure of the courts according to whether the charge was felony or misdemeanour, and other matters that depended on the distinction included the power of the police to arrest a suspect on suspicion that he had committed an offense (which was generally permissible in felony, but not in misdemeanour). By the early 19th century it had become clear that the growth of the law had rendered this classification obsolete and in many cases inconsistent with the gravity of the offenses concerned (theft was a felony, irrespective of the amount stolen; obtaining by fraud was always a misdemeanour). Efforts to abolish the distinction in English law did not succeed until 1967, when the distinction was replaced by that between arrestable offenses and other offenses (an arrestable offense is one punishable with five years' imprisonment or more; offenders may be arrested for other offenses subject to certain conditions). In later legislation it has proved necessary to devise further classifications--for the purposes of powers of investigation, a category of serious arrestable offenses has been created, and, for the purpose of deciding in which courts the case should be tried, a different classification of offenses into three categories of indictable, "either way," and summary has been adopted. The traditional classification between felony and misdemeanour has been retained in many U.S. jurisdictions (although there has been rationalization of the allocation of offenses to one category or the other) and is used as the basis of determining the

court that will hear the case. In some jurisdictions a further class of offense (violations) has been added, to include minor offenses (corresponding broadly to the English category of summary offenses).

SOME PARTICULAR CRIMES

Murder.

In English tradition, murder was defined as the willful killing with malice aforethought of a human creature in being within the king's peace, the death occurring within a year and a day of the injury. Most of these elements remain in modern definitions of the crime--the requirement that the victim is "in being," for instance, distinguishes abortion from murder--although in some respects the definition has become more complex. Many of the problems of defining murder have centred on the mental element--the "malice aforethought." The old English rule extended this concept to include not only intentional or deliberate killings but also accidental killings in the course of some other serious crime (such as robbery or rape). This rule, the felony murder rule, was adopted in many other jurisdictions, although it has often produced harsh results when death has been caused accidentally in the course of what was intended to be a minor crime. The rule was abolished in England in 1957, but since then English law has been in a state of confusion over the precise definition of murder. It is now settled that an intention to kill is not necessary and that an intention to cause serious bodily injury is sufficient, but the precise interpretation of intention in this context remains controversial. Similar problems have arisen in many U.S. jurisdictions, some of which distinguish between different degrees of murder--first-degree murder may require proof of premeditation over and above the normal requirement of intention. Surprisingly, murder and manslaughter are not mentioned in the Qur`an and are subject in Islamic countries to customary law as amended by Shari'ah. Virtually all systems treat murder as a crime of the utmost gravity, providing in some cases the death penalty or a special form of sentence, such as a mandatory life sentence, often with restrictions on parole. A high proportion of murders in all societies are committed spontaneously by persons acquainted with the deceased, often a

member of the same family, as a result of quarrels or provocation. The convicted murderer is often a person with no other criminal conviction.

Rape.

The traditional legal definition of rape is the performance of sexual intercourse by a man other than her husband with a woman against her will, by force or fraud. This definition has been adapted in the statutes of some jurisdictions; in Canada the crime of rape has been abolished as a separate offense and merged into a wider general category of sexual assault. Most jurisdictions do not treat as rape an act of sexual intercourse by a husband with his wife without her consent, unless the marriage has effectively been terminated by a legally recognized separation. Although many rapes involve the application or threat of violence, it is possible to commit rape by fraud--either by persuading the victim that what is to take place is not sexual intercourse (by representing it as medical treatment, for instance) or by impersonating some other person, such as the victim's husband. Under the provisions of most criminal codes, rape requires penetration of the female organ by the male organ (but does not require ejaculation); other forms of sexual abuse (such as oral penetration or anal penetration) are dealt with, if at all, under different provisions. In the United States, however, the crime of rape may also include forcible sodomy, and victims of rape may be women or men. In many rape trials the issue is whether the victim consented to the sexual intercourse, and this may lead to distressing cross-examination, in some cases about the woman's previous sexual behaviour, whether with the accused or with other persons. In many jurisdictions cross-examination of the complainant on such matters is now restricted, and the embarrassment of the complainant is further mitigated by provisions restricting publication of the woman's identity. Proof is made more difficult by the common need to prove not only that the victim did not consent but also that the accused knew this--or at least was aware of that possibility.

When guilt is established, rape in most systems of criminal law is treated as a grave crime; 95 percent of those convicted of rape in England, for instance, are sentenced to imprisonment. A high proportion of rapists escape conviction for a

variety of reasons. The victim may be reluctant to report the incident, possibly because of fears of hostile treatment by investigating authorities or by defense lawyers in court; there is a higher than average acquittal rate of those indicted for rape, as a result of the difficulty of proving a crime of which there are rarely any witnesses other than the complainant and the accused. The motivation of rapists is now acknowledged to be a more complex matter than was formerly believed; it has come to be widely accepted that rape is not necessarily the result of sexual desire but is more likely to be motivated by aggression and the desire to humiliate or exercise domination over the victim.

Incest.

The crime of incest consists of sexual intercourse between near relatives. Incest was not a crime under the common law but was punishable historically in the ecclesiastical courts. Legislation prohibiting incest was enacted in England in 1908, and most English-speaking jurisdictions now prohibit intercourse between close relatives, but there are differences among the systems in the relationships within which intercourse is forbidden. Most systems forbid intercourse between immediate relatives--father and daughter, brother and sister, mother and son. There are some anomalies--English law prohibits intercourse between grandfather and granddaughter but not between grandmother and grandson. Consent to intercourse is irrelevant to the charge of incest, but if there is no consent, the crime of rape may also be committed. Generally, sexual intercourse between family members who are not related by blood--for instance, stepfather and stepdaughter--is not considered incest, but this is prohibited in some jurisdictions. Both parties are considered guilty if incest occurs, but in many systems there is an exemption from liability for women below a certain age (16 in England, 18 in some U.S. jurisdictions). Most cases of incest that come before criminal courts concern sexual intercourse between fathers and relatively young daughters, and it is believed that incest in this form is far more common than the statistics of court cases suggest. Treatment of the offenders in such cases presents acute difficulties to the courts--on the one hand, the offense is widely regarded as serious, involving sexual abuse

of children and a breach of the parent's responsibility for the child's welfare; on the other hand, to impose a severe penalty, such as imprisonment, on the father may result in the destruction of the family unit and the infliction of other deprivations on the child victim, in particular feelings of guilt for being responsible for the imprisonment of the father.

Perjury.

Perjury originally consisted in the giving of false evidence on oath to a court of law; it has now been expanded in most jurisdictions to include evidence given otherwise than on oath (under affirmation, for instance, by a person who objects to swearing) and to other tribunals that have the authority of the law. Perjury may be committed by witnesses from either the prosecution or the defense (or by witnesses on either side in civil litigation) and in proceedings before the jury or after the verdict in proceedings leading to sentence. As a general rule, the accused person must make a false statement that he either knows to be false or does not believe to be true, and the false statement must normally be material to the matters in issue in the proceedings. In many jurisdictions the law imposes special requirements for the proof of perjury--it is not normally sufficient to rely on the evidence of one witness as to the falsity of the alleged perjured statement. Crimes associated with perjury include subornation of perjury (persuading other persons to commit perjury) and a wide variety of statutory offenses involving making false statements in official documents (such as applications for drivers' licenses) that are not normally treated so seriously.

Prostitution.

The laws regulating prostitution vary greatly from one jurisdiction to another. In some jurisdictions prostitution--generally defined as the provision of sexual services for money--is itself illegal; in others the act of prostitution is not illegal in itself, but many associated activities are unlawful. English law, for instance, does not prohibit prostitution itself but does prohibit soliciting for prostitution in a public place, living on the earnings of prostitution, exercising control over prostitutes, or keeping a brothel (any premises where two or more prostitutes are

employed). In some jurisdictions, notably in the U.S. state of Nevada, prostitution is lawful and practiced openly subject only to health and related controls.

Arson.

In common law, arson consisted of setting fire to the dwelling of another person. Subsequent statutes have expanded the scope of the offense to include setting fire to other types of buildings, and in English law any kind of damage deliberately caused by fire--even setting fire to rubbish--is now arson, but generally setting fire to a building is necessary. The gravity of the crime may depend on the extent to which life is endangered--the law may distinguish between arson endangering life, or arson of occupied buildings, and other forms of arson, but most systems consider the crime a serious one. The motivation of those who commit arson differs--arson may be committed as an act of revenge against an employer or by a jealous lover, for example, or by persons who find excitement in fires or have pathological impulses to set fires. Schools are sometimes set on fire by pupils out of resentment or simple vandalism. Some arson is more rationally motivated--a burglar may set fire to a house to conceal the evidence of his crime, as may an employee who is anxious to conceal accounts from an auditor. Another phenomenon is setting fire to premises belonging to the fire setter in order to make a fraudulent insurance claim.

Theft and burglary.

Theft, sometimes still known by the traditional name of larceny, is probably the most common crime involving a criminal intent. The crime of grand larceny in some U.S. jurisdictions consists of stealing more than a specified sum of money or property worth more than a specified amount. The traditional definition of theft specified the physical removal of an object that was capable of being stolen, without the consent of the owner and with the intention of depriving the owner of it permanently. This intention, which has always been an essential feature of theft, does not necessarily mean that the thief must intend to keep the property--an intention to destroy it, or to abandon it in circumstances where it will not be found,

is sufficient. In many legal systems the old definition has been found to be inadequate to deal with modern forms of property that may not be physical or tangible (a bank balance, for instance, or data stored on a computer), and more sophisticated definitions of theft have been adopted in modern legislation. The distinction that the common law made between theft (taking without consent) and fraud (obtaining with consent, as a result of deception) has been preserved in many modern statutes, but the two crimes are rarely regarded as mutually exclusive, as they were in the past. It is now accepted that an act may constitute both theft and fraud, as in the theft and subsequent sale of an automobile.

Burglary consisted originally of breaking into a dwelling by night with intent to commit a felony, but as in the case of many other crimes the definition has been expanded in many legal systems. In English law, any entry by an individual into a building as a trespasser with intent to commit theft or certain other offenses is burglary, and some jurisdictions recognize an offense of burglary of an automobile—breaking into it to steal the contents. The essence of burglary is normally the entry into a building with a criminal intent. Entry without the intent to commit a crime of the kind specified in the burglary statute is not burglary--it is merely a trespass, which is not criminal in many jurisdictions. Although the motivation of most burglars is theft, an intention to commit various other offenses converts a trespass into a burglary--it is possible, for instance, to commit burglary with intent to rape.

Robbery is the commission of theft in circumstances of violence. It involves the application or the threat of force in order to commit the theft or to secure escape. Robbery takes many forms--from the mugging of a stranger in the street, in the hope of stealing whatever he may happen to have in his possession, to much more sophisticated robberies of banks or similar premises, involving numerous participants and careful planning.

OTHER MODES OF CRIMINAL ACTIVITY

Organized crime.

In addition to that segment of the population made up of individual criminals acting independently or in small groups, there exists a so-called underworld of criminal organizations engaged in offenses such as cargo theft, fraud, robbery, kidnapping for ransom, and the demanding of "protection" payments. In the United States and Canada, the principal source of income for organized crime is the supply of goods and services that are illegal but for which there is continued public demand. Examples include drugs, prostitution, loan-sharking (lending money at extremely high rates of interest), and gambling. To ensure freedom from the law, the organizations must subvert both the police and the courts.

Organized crime in the United States is best viewed as a set of shifting coalitions, normally local or regional in scope, between groups of gangsters, business people, politicians, and union leaders. Many of these people have legitimate jobs and sources of income. So-called street-level criminals are normally independent of major crime syndicates. Among other advanced industrial nations, the closest similarities to this organizational model occur in Australia, where extensive narcotics, cargo theft, and labour racketeering rings have been discovered, and in Japan, where there are gangs that specialize in vice and extortion. In Britain groups of organized criminals have not developed in this way, principally because the supply and consumption of alcohol and opiates, gambling, and prostitution remain legal but partly regulated, owing to a more liberal and pragmatic attitude of successive governments aware of the impossibility of total enforcement. This apparent laxity reduces the profitability of supplying such demands criminally. Far Eastern groups such as the Chinese Triads are important in the supply of drugs to and by way of Britain. Except for cargo thieves who work at airports and local vice, protection, and pornography syndicates, British crime organizations tend to be relatively short-term groups drawn together for specific projects, such as fraud and armed robbery, from a pool of long-term professional criminals.

In many Third World countries, apart from the drug trade, the principal form of organized crime is black-marketeering, including smuggling and corruption in the granting of licenses to import goods and to export foreign exchange. Armed

robbery, cattle theft, and maritime piracy and fraud are organized crime activities in which politicians have less complicity. Robbery is particularly popular and easy because of the widespread availability of arms supplied to nationalist movements by those who seek political destabilization of their own or other states, and who may therefore exploit the dissatisfaction of ethnic and tribal groups.

"White-collar" crime.

Crimes committed by business people, professionals, and politicians in the course of their occupation are known as "white-collar" crimes, after the typical attire of their perpetrators. Contrary to popular usage, criminologists tend to restrict the term to those illegal actions intended by the perpetrators principally to further the aims of their organizations rather than to make money for themselves personally. Examples include conspiring with other corporations to fix prices of goods or services in order to make artificially high profits or to drive a particular competitor out of the market; bribing officials or falsifying reports of tests on pharmaceutical products to obtain manufacturing licenses; and constructing buildings or roads with cheap, defective materials while charging for components meeting full specifications. Often such activities are attributable to over-enthusiastic employees or executives acting on their own initiative, but sometimes they represent a form of "upperworld" organized crime.

The cost of corporate crime in the United States has been estimated at $200,000,000,000 a year--three times the cost of organized crime. Such crimes have a huge impact upon the safety of workers, consumers, and the environment, but they are seldom detected. Compared with crimes committed by juveniles or the poor, corporate crimes are very rarely prosecuted in the criminal courts, and executives seldom go to jail, though companies may pay large fines.

The term white-collar crime is used in another sense, by the public and academics, to describe fraud and embezzlement. Rather than being crime "by the firm, for the firm," this constitutes crime for profit by the individual against the organization, the public, or the government. (Tax fraud costs at least 5 percent of the gross national product in most developed countries.) Because of the concealed nature of

many frauds and the fact that few are reported even when discovered, the cost is impossible to estimate precisely, but in the United States it is thought to be at least 10 times the combined cost of theft, burglary, and robbery.

Terrorism.

From the 1960s, international terrorist crimes, such as the hijacking of passenger aircraft, political assassinations and kidnappings, and urban bombings, constituted a growing phenomenon of increasing concern, especially to Western governments. Most terrorist groups are associated either with millenarian revolutionary movements on an international scale (such as some Marxist organizations) or with nationalist movements of particular ethnic, religious, or other cultural focus.

Three broad categories of terrorist crime may be distinguished, not in legal terms, but by intention. Foremost is the use of violence and the threat of violence to create public fear. This may be done by making random attacks to injure or kill anyone who happens to be in the vicinity when an attack takes place. Because such crimes deny, by virtue of their being directed at innocent bystanders, the unique worth of the individual, terrorism is said to be a form of crime that runs counter to all morality and so undermines the foundations of civilization. Another tactic generating fear is the abduction and assassination of heads of state and members of governments in order to make others afraid of taking positions of leadership and so to spread a sense of insecurity. Persons in responsible positions may be abducted or assassinated on the grounds that they are "representatives" of some institution or system to which their assailants are opposed.

A second category of terrorist crime is actual rule by terror. It is common practice for leaders of terrorist organizations to enforce obedience and discipline by terrorizing their own members. A community whose collective interests the terrorist organization claims to serve may be terrorized so that their cooperation, loyalty, and support are ensured. Groups that come to power by this means usually continue to rule by terror.

Third, crimes are committed by terrorist organizations in order to gain the means for their own support. Bank robbery, kidnapping for ransom, extortion, gambling

rake-offs (profit skimming), illegal arms dealing, and drug trafficking are among the principal crimes of this nature. In the Middle East, hostages are frequently sold as capital assets by one terrorist group to another.

Measurement of crime

Estimating the amount of crime actually committed has troubled criminologists for many years; the figures for recorded crime do not give an accurate picture because they are influenced by variable factors such as the willingness of victims to report crimes. It is widely believed that only a small fraction of the crime actually committed is reported to authorities. For this reason the criminal who is detected is not necessarily representative of all those who commit crime, and thus attempts to explain the causes of crime by reference to those who are identified as criminals must be approached with caution.

The public's view of the frequency and gravity of crime, obtained largely from the news media, may be seriously distorted, as the media tend to concentrate on serious or sensational crimes and often fail to give a full and accurate picture of what has happened. A brief report of a case in court, for instance, inevitably is selective; much of the evidence that the court has heard is omitted. A more detached view may be provided by detailed statistics of crime compiled and published by a department of government--in the United States, for instance, the Federal Bureau of Investigation publishes an annual, the Uniform Crime Reports, and in England the Home Office produces each year a volume entitled Criminal Statistics, England and Wales, which (among other things) gives an account of the trends in different types of crime. Official statistics such as these are frequently used by policymakers as the basis for new procedures in crime control--they may show, for instance, that there has been an increase in the incidence of a particular type of crime over a period of years and suggest, therefore, that some change in the methods of dealing with that type of crime is necessary. In fact, many official statistics of crime are subject to serious error and may be almost as misleading as the general impressions formed by the public through the news media, particularly if they are used without an understanding of the processes by which they are

compiled and the limitations to which they are necessarily subject. The statistics are usually compiled on the basis of reports from police forces and other law enforcement agencies and are generally known as statistics of reported crime, or crimes known to the police. Because only incidents observed by the police or reported to them by victims or witnesses are included in the reports, the picture of the amount of crime actually committed may be distorted. One factor accounting for this distortion is the extent to which police resources are allocated to the investigation of one kind of crime rather than another, particularly with regard to what are known as "victimless crimes," such as possession of drugs. These crimes are not discovered unless the police set out to look for them, and they do not figure in the statistics of reported crime unless the police take the initiative; thus, a sudden increase in the reported incidence of a crime from one year to the next may merely show that the police have taken more interest in that crime and devoted more resources to its investigation. Ironically, efforts to discourage or eliminate a particular kind of crime through more vigorous law enforcement may create the impression that the crime concerned has increased rather than decreased, because more instances are detected and thus enter the statistics.

A second factor that can have a striking effect on the apparent statistical incidence of a particular kind of crime is a change in the willingness of victims of the crime to report it to the police. It is believed by most criminologists, on the basis of research, that the crime reported to the police amounts to only a small proportion of the crime actually committed. Estimates of the so-called dark figure (the number of unreported crimes) vary, but it is thought that in some cases the reported crimes may constitute less than 10 percent of those actually committed. Victims of crimes have many reasons for not reporting them: they may not realize that a crime has been committed against them (children who have been sexually molested, for instance); they may believe that the police will not be able to detect the offender; they may be afraid of involvement in the processes of the law as witnesses; they may be embarrassed by their own conduct that has led them to become the victim of the crime (a man robbed by a prostitute, for instance, or a person who has been

the victim of a confidence trick as a result of his own greed or credulity). A particular type of crime may not appear sufficiently serious to make it worthwhile to inform the police, or there may be ways in which the matter can be resolved without involving them--an act of violence by one schoolchild against another may be dealt with by the school authorities, or a dishonest employee may be dismissed without prosecution. All of these factors are difficult to measure with any degree of accuracy, and there is no reason to suppose that they remain constant over a period of time. Thus, a change in any one of these factors may produce the appearance of an increase or a decrease in a particular kind of crime, when in fact there has been no such change, or the real change has been on a much smaller scale than the statistics suggest.

A third factor that may affect the picture of crime presented by official statistics is the way in which the police treat particular incidents. Many of the laws defining crimes are imprecise or ambiguous—concepts such as reckless driving, obscenity, and gross negligence may leave a great deal to interpretation. The result may be that conduct which is treated as a crime in one police district, and thus appears as such in statistics, may not be treated as the same crime in another, because the law is interpreted in a different manner. Another practice that may have the same result is the way in which a particular incident is broken down into different crimes. The theft of a number of items may be recorded as a single theft of all of them or as a series of thefts of the individual items.

Criminologists have for many years endeavoured to obtain a more accurate picture of the incidence of crimes and the trends and variations from one period to another. Two research methods have usually been employed--the victim survey and the self-report study. The victim survey requires the researcher to identify a sample of the population at risk of becoming victims of the kind of crime in which the researcher is interested, or of crimes generally, and to ask them to disclose any crime of which they have been victims during the period specified in the research. The information obtained from the survey, after a large number of people have been questioned, is compared with the statistics for reported crime for the same

period and locality, giving an indication of the relationship between the actual incidence of the type of crime in question and the number of cases of that type reported to the police. Although criminologists have developed sophisticated procedures for interviewing victim populations, such projects are subject to a number of limitations. Results depend entirely on the victim's recollection of the incidents, his ability to recognize that a crime has been committed, and his willingness to disclose it. The method can be applied only to crimes that have victims; it does not help to identify the incidence of victimless crimes. Research of this kind, however, undertaken with an awareness of the limitations of the process, undoubtedly extends knowledge of the actual incidence of particular types of crime and, if it is carried on over a period of years, may provide a clearer picture of trends in crime than official statistics. One major survey of this kind, the British Crime Survey, is expected to last for many years and, as it obtains information from a very large sample of households, may be more representative than smaller ones.

An alternative approach favoured by some criminologists is the self-report study, in which a sample of the general population is asked, under assurances of confidentiality, if they have committed any offenses of a particular kind. This type of research is subject to the same difficulty as the victim survey--the researcher has no means of verifying the information given to him, and the subject can easily conceal the fact that he has committed an offense at some time--but surveys of this kind have often confirmed that large numbers of offenses have been committed without being reported and that crime is much more widespread than official statistics suggest.

Analysis of crime

CHARACTERISTICS OF OFFENDERS

Gender patterns.

Knowledge of the types of people who commit crimes is subject to one overriding limitation: it is generally based on studies of those who have been detected, prosecuted, and convicted. The populations of penal institutions are not necessarily

representative of the whole range of criminals--in one sense, they are by definition the unsuccessful criminals. Despite this limitation, some basic facts emerge that probably give a reasonably accurate picture of those who commit crimes. The first is that crime is predominantly a male activity. In all criminal populations, whether of offenders passing through the courts or of those sentenced to institutions, men outnumber women by a high proportion. In Britain in 1984, for instance, of 449,000 offenders found guilty of criminal offenses, 387,400 (86 percent) were males; in the same year, the daily average population of the prisons consisted of 41,822 men and 1,473 women. In most Western societies the incidence of recorded crime by women, and the number of women passing through the penal systems, is on the increase; in the United States, for instance, the number of women arrested for property crimes between 1960 and 1976 increased by 276 percent--a significantly higher rate of increase than that exhibited by other groups. A similar trend is shown in English prison statistics: the number of women in prison under sentence rose from 538 in 1974 to 941 in 1984, an increase of 75 percent in 10 years. A number of explanations have been offered for this trend. One suggestion is that it reflects a real trend in the commission of crimes by women--that the changing social role of women, with more women leaving the home and taking employment, expecting and achieving financial independence, leads to greater opportunity for crime and to greater temptation. An alternative explanation is that the change in the apparent rate of female criminality merely reflects a change in the operation of the criminal justice system--that crimes committed by women are less likely than was previously the case to be ignored by law enforcement agencies out of a sense of chivalry. Even though female criminality appears to be increasing faster than male criminality, it will be many years before women reach the same level of crime as men.

Age patterns.

A second aspect of criminality about which there is a reasonable measure of agreement is that crime is predominantly an activity of the young. In both Britain and the United States, for example, the peak period for involvement in relatively

minor property crime is adolescence--from 15 to 21. For involvement in more serious crimes the peak age is likely to be rather higher, from the late teenage years through the 20s. Criminality tends to decline steadily after the age of 30. Criminologists have sought explanations of this phenomenon--whether it is a natural effect of aging, the consequence of taking on family responsibilities, or the effect of experiencing penal measures imposed by the law for successive convictions--but the evidence is inconclusive. Not all types of crime are subject to decline with aging. Fraud and certain kinds of theft, as well as crimes requiring a high level of businesslike organization, are more likely to be committed by older men, and sudden crimes of violence, committed for emotional reasons, may occur at any age.

The relationship between social class or economic status and crime has been studied extensively by criminologists. Studies carried out in the United States in the 1920s and '30s claimed to show that a higher incidence of criminality was concentrated in deprived and deteriorating neighbourhoods of large cities, and studies of penal populations revealed that the level of educational and occupational attainments was generally lower than in the wider population. Early studies of juvenile delinquents dealt with by courts disclosed a high proportion of lower-class offenders. Later research has called into question the assumption that criminality is closely associated with social origin; in particular, self-report studies have suggested that offenses are more widespread across the social spectrum than the figures based on identified criminals would suggest.

The relationship between racial or ethnic origin and criminality is a difficult and controversial question. Penal populations probably contain a disproportionately high number of persons from some minority racial groups, in the sense that the proportion of minority group members in prison is greater than the group's proportion in the general population. Criminologists have pointed out that this may be the result of the high incidence among minority racial groups of characteristics that are commonly associated with identified criminality--e.g., unemployment and low economic status--and the fact that in many cities racial minority groups inhabit

areas that have traditionally been high crime areas, perhaps as a result of their shifting populations and general lack of social cohesion. Further explanations are differential enforcement practices on the part of the police and the adherence of members of some minority groups to cultural standards that are in conflict with the general law (e.g., the widespread use of cannabis [marijuana] by members of the predominantly black Rastafarian sect).

THEORIES OF CAUSATION

Few modern criminologists would claim that any single theory constitutes a universal explanation of criminality or a valid predictor of future criminal behaviour in a particular population. A more common view is that many of the different theories offered may help to explain particular aspects of criminality and that different types of explanation may all contribute to the understanding of the problem of crime.

Biological theories.

Some theories attribute the tendency toward criminality to innate biological factors. The most famous of these is probably that of the Italian Cesare Lombroso (1835-1909), one of the first scientific criminologists, whose theories were related to Darwinian theories of evolution. His investigations of the skulls and facial features of robbers led him to the hypothesis that serious or persistent criminality was associated with atavism, or the reversion to a primitive stage of human development. Another biological theory related criminality to body types, suggesting that it was more common among muscular, athletic persons (mesomorphs) than among tall, thin persons (ectomorphs) or soft, rounded individuals (endomorphs). These theories have little support today, but there is some interest in the idea that criminality may be related to chromosomal abnormalities--in particular, the idea that so-called XYY males (characterized by the presence of a surplus Y chromosome) may be more likely to be involved in criminal behaviour than the general population.

Some criminologists have endeavoured to answer the question of whether biological factors are more important than social factors in criminal behaviour by

studying the behaviour of twins. Various studies have shown that twins are more likely to exhibit similar tendencies toward criminality if they are identical (monozygotic) than if they are fraternal (dizygotic). The suggestion of genetic influences in criminal behaviour is supported by studies of adopted children carried out to determine the influence of the biological parent on criminality. One such study showed that the rate of criminality was higher among those adopted children who had one biological parent who was criminal than among those who had one adoptive parent who was criminal but whose biological parents were not. The highest rates of criminality were found among those children who had both biological parents and adoptive parents who were criminal.

Sociological theories.

Sociologists have proposed a variety of theories that explain criminal behaviour as a normal adaptation to the offender's social environment. One such theory, known as differential association, proposed that all criminal behaviour is learned behaviour and that the process of learning criminal behaviour depends on the extent of the individual's contact with other persons whose behaviour reflects varying standards of legality and morality. The more the individual is exposed to contact with persons whose own behaviour is unlawful, the more likely he is to learn and adopt their values as the basis for his own behaviour. The theory of anomie, proposed by the American Robert K. Merton, suggested that criminality is a result of the offender's inability to attain by socially acceptable means the goals that society expects of him; faced with this inability, the individual is likely to turn to other, not necessarily socially acceptable, objectives or to pursue the original objectives by unacceptable means. A development from this theory is the concept of the subculture--an alternative set of moral values and conventional expectations to which the person can turn if he cannot find acceptable routes to the objectives held out for him by the broader society. This theory, developed particularly with reference to delinquent gangs in U.S. cities, has been disputed by other sociologists who deny the existence of any subculture of delinquency among the lower classes

of society; the behaviour of gangs is for these latter sociologists an expression of widespread lower-class values emphasizing toughness and excitement.

A further group of sociological theories denies the existence of subcultural value systems and portrays the delinquent as an individual who subscribes generally to the morals of society but who is able to justify to himself particular forms of delinquent behaviour by a process of "neutralization," in which the behaviour is redefined in moral terms to make it acceptable. Control theory emphasizes the links between the offender and his social group--the individual's bond to society. According to this theory, the ability of the individual to resist the inclination to commit crime--which may be an easy way to satisfy a particular desire--depends on the strength of his attachment to parents, his involvement with conventional activities and avenues of progress, and his commitment to orthodox moral values that prohibit the conduct in question. Labeling theory, by contrast, portrays criminality as a product of the reaction of society to the individual, rather than of his own inclinations and personality. It assumes that the criminal is not

substantially different from any other individual, except that he has become involved in the processes of the criminal justice system and has acquired a "criminal" identity. Through a process of rejection by law-abiding persons and acceptance by other delinquents, which is a consequence of the criminal identity conferred on him by the courts, the offender becomes more and more socialized into criminal behaviour patterns and estranged from law-abiding behaviour. Eventually he comes to see himself cast by society into the role of a criminal, and he acts out society's expectations. Each time he passes through the court system, the process is extended to form a process described as "amplification of deviance." Radical criminologists change the focus of inquiry, looking for the causes of delinquency not in the individual but in the structure of society, in particular its political and legal systems. The criminal law is seen as an instrument by which the powerful and affluent maintain their position and coerce the poor into patterns of behaviour that preserve the status quo.

Psychological theories.

Psychologists have approached the task of explaining delinquent behaviour by examining in particular the processes by which behaviour and restraints on behaviour are learned. Psychoanalytical theories emphasize the instinctual drives for gratification and the control exercised through the more rational aspect of personality, the superego. Criminality is seen to result from the failure of the superego, as a consequence either of its incomplete development or of unusually strong instinctual drives. The empirical basis for such a theory is necessarily thin. Behaviour theory views all behaviour--criminal and otherwise--as learned and thus manipulable by the use of reinforcement and punishment. Social learning theory examines the manner in which behaviour is learned from contacts within the family and other intimate groups, from social contacts outside the family, particularly from peer groups, and from exposure to models of behaviour in the media, particularly television.

Mental illness is the cause of a relatively small proportion of crimes, but its importance as a causative factor may be exaggerated by the seriousness of some of the crimes committed by persons with mental disorders. Severe depression or psychopathy (sometimes described as sociopathy or personality disorder) may lead to grave offenses of violence. On a less serious level, depression may lead to theft or other uncharacteristic behaviour.

A non-Western perspective: China.

The Chinese have in general adopted a Marxist interpretation of the causes of crime. Crime is viewed as a product of class society, of exploitative systems founded upon the institution of private property. Because the socialist system is considered by its proponents as incapable of producing crime, official theory has always looked outside of post-1949 Chinese society to find the causes of contemporary crime. A number of specific sources of criminal activity have been suggested: (1) external enemies and remnants of the overthrown reactionary classes (the latter referring to the government of the Republic of China in Taiwan) who infiltrate the country with spies and conduct sabotage; (2) remains of the old (pre-1949) society, such as gangsters and hooligans, who refuse to reform; (3)

lingering aspects of bourgeois ideology that prize profit, cunning, selfishness, and decadence and thus encourage crime; and (4) the poverty and cultural backwardness that is seen as the legacy of the old society. The Cultural Revolution (1966-76) has also been cited as a cause of crime; it is said to have confused notions of right and wrong and to have destroyed respect for authority.

While Chinese criminology thus adopts a social explanation of crime in capitalist society, it has little sympathy for the view that society is to blame for crime in contemporary China. The two main causes are seen to be backward thinking and ignorance. For this reason, crime is ideally to be fought, and ultimately eliminated, by thought reform and by education.

Detection of crime

In most countries the detection of crime is the responsibility of the police, although special law enforcement agencies may be responsible for the discovery of particular types of crime (customs departments, for instance, may be responsible for the detection of smuggling and related offenses). Crime detection falls into three distinguishable phases: the discovery that a crime has been committed, the identification of a suspect, and the collection of sufficient evidence to indict the suspect before the court. Criminologists have shown that a high proportion of crimes are discovered and reported by persons other than the police (such as victims or witnesses), but certain types--in particular crimes that may involve a subject's assent, such as dealing in drugs or prostitution, or those in which there may be no identifiable victim, such as obscenity--are often not discovered unless the police take active steps to determine whether these crimes are being committed. This may require controversial methods, such as surveillance, interception of communications, infiltration of gangs, and entrapment (e.g., by making a purchase from a suspected drug dealer). Once the commission of a crime has been discovered, the identification of the suspect becomes essential.

THE ROLE OF FORENSIC SCIENCE

Forensic science has come to play an increasingly important part in the investigation of serious crimes. One of the first significant developments was

identification by fingerprints. It was discovered in the 19th century that almost any contact between a finger and a fixed surface left a latent mark that could be exposed by a variety of procedures, the most common being the use of a fine powder. It was accepted in 1893, by the Troup Committee established by the Home Secretary, that no two individuals had the same fingerprints, and this proposition has never been seriously refuted. Fingerprint evidence was accepted for the first time in an English court in 1902. The original purpose of recording and collecting fingerprints was to establish and to make readily available the criminal record of particular offenders, but fingerprinting is now widely used as a means of identifying the perpetrators of particular offenses. Most major police forces maintain collections of fingerprints taken from known criminals at the time of their conviction, for use in identifying these individuals should they commit later crimes. Fingerprints (which may be incomplete) found at the scene of the crime are matched with fingerprints in the collection. According to the British standard, if the sets of fingerprints share at least 16 characteristics, it is considered virtually certain that they are from the same person. Searching fingerprint collections had historically been a time-consuming manual task, based on various systems of classification, but systems for electronic storage and rapid searching of fingerprint collections were developed and implemented in the 1980s.

A broad range of other scientific techniques is available to law enforcement agencies attempting to identify suspects or to establish beyond doubt the connection between a suspect and the crime in question. Examples include the analysis of bloodstains and traces of other body fluids (such as semen or spittle) that may indicate some of the characteristics of the offender. Fibres can be analyzed by microscopy or chemical analysis to show, for instance, that fibres found on the victim or at the scene of the crime are similar to those in the clothing of the suspect. Hair samples, and particularly skin cells attached to hair roots, can be compared chemically and genetically to those of the suspect. Many inorganic substances, such as glass, paper, and paint, can yield considerable information under microscopic or chemical analysis. Examination of a document in question

may reveal it to be a forgery, on the evidence that the paper on which it is written was manufactured by a technique not available at the time to which it allegedly dates. The refractive index of even small particles of glass may be measured to show that a given item or fragment of glass was part of a particular batch manufactured at a particular time and place. Such information may help to identify the kind of automobile involved in a hit-and-run accident. Computer networks allow investigators to search increasingly large bodies of data on material samples, but the creation of the necessary data bases is a lengthy process.

MODUS OPERANDI AND SUSPECT IDENTIFICATION

The method by which an offense was committed may also help to identify the suspect, as many offenders repeatedly commit offenses in much the same way. The burglar's method of entry into the house, the type of property stolen, or the kind of deception practiced on the victim of a fraud may all suggest to the police who is responsible for the crime. Visual identification of a stranger by the victim is often possible, but experience has shown that such identifications are often mistaken and have frequently led to miscarriages of justice. If the victim or witness believes that he can recognize the offender, the police may show him an album containing photographs of a large number of known criminals, in the hope that one can be picked out. A suspect identified in this way is usually asked to take part in a lineup, in which the witness is asked to pick the suspect out of a group of people with similar characteristics.

GATHERING EVIDENCE

The identification of the suspect is not the final stage of the process: it is essential that the investigating agency gather sufficient legally admissible evidence to convince the judge or jury that the suspect is guilty before a conviction can be expected. It is common for the police to be reasonably certain that a particular individual is responsible for a crime but to remain unable to establish his guilt by legally admissible evidence. In order to secure the necessary evidence, the police employ a variety of powers and procedures; because these potentially involve interference with the freedom of the suspect (who must at this stage be treated as

an innocent person), they are normally subject to close control either by legislation or by the courts.

One important procedure is a search of the person of the suspect or of premises or vehicles. Most jurisdictions within the common-law tradition allow a search to be carried out only if there is "probable cause for believing" or "reasonable ground for suspecting" that the evidence will be found. In some cases a person may be stopped on the street and searched, subject to various requirements that the police officer identify himself and state the reasons for the search. A search of private premises usually requires a search warrant issued by a magistrate or judge. The law generally permits a search warrant to be issued only if the issuing authority is satisfied after hearing evidence on oath that there is good reason to suspect that the evidence, which the warrant usually defines specifically, will be found on the premises. The warrant may be subject to time limits and normally permits only one search to be carried out. In most countries the judge or magistrate who issues the warrant must be told of the outcome of the search. Material seized as a result of a search under the authority of a search warrant is usually detained by the police for production as exhibits at any subsequent trial. In the United States the law has imposed strict consequences on any abuse of this procedure; evidence discovered as a result of any search that does not comply with the procedures and standards laid down by the Supreme Court and by other courts, interpreting the various amendments to the U.S. Constitution collectively known as the Bill of Rights, is not admitted in the trial, even though it clearly establishes the guilt of the accused person, and even though the suppression of the evidence may prevent the conviction of a person who is plainly guilty. This rule, known as the exclusionary rule, has given rise to controversy in the United States and has not generally been adopted in other English-speaking countries.

INTERROGATION AND CONFESSION

Miranda warnings.

The interrogation of suspected persons is an important aspect of the investigation of offenses. Usually the aim of the questioning is to obtain an admission of the

offense that will lead eventually to a plea of guilty and avoid the need for a contested trial. All English-language countries place restrictions on the scope and methods of interrogation in order to ensure that suspects are not coerced into confessions by unacceptable means. In the United States any suspect who is being interrogated in custody must be offered the services of a lawyer, at the expense of the state if he cannot afford to pay, and failure to advise the suspect of this right (known as the Miranda warnings, after the case of Miranda v. Arizona) results in the rejection of a confession as evidence.

The Judges' Rules.

English law follows the same general principle, that a person suspected or accused of a criminal offense is not at any stage in the process of investigation or trial obliged to answer any question or to give evidence. (There are a few minor exceptions; for instance, the owner of a motor vehicle is required by law to disclose the identity of the person who was driving the vehicle on any particular occasion, and drivers of motor vehicles may be required to give samples of breath, blood, or urine in certain circumstances.) For many years the law relating to confessions in England consisted of a simple rule prohibiting the admission as evidence at trial of any involuntary statement made by an accused person. This rule was supplemented by more detailed rules governing the questioning of suspected persons by the police, formulated by the judges of the High Court and known as the Judges' Rules. The principal effect of the Judges' Rules was to impose an obligation on the investigating police officer to administer to the suspect a caution to the effect that he was not obliged to answer any question and that anything he did say might be given in evidence at his trial. This caution was required to be given at the beginning of any period of interrogation and immediately before the suspect began to make a full statement or confession. Failure to give the caution at the right time or in the right form did not necessarily mean that the statement would be excluded from evidence, but it did give the trial judge the discretion to exclude the evidence if he considered it just to do so. The operation of the Judges' Rules was a source of controversy for many years, and they have been replaced by

a comprehensive series of provisions under the Police and Criminal Evidence Act, 1984. This act provides that a confession by an accused person may be admitted in evidence provided that the court is satisfied that the confession was not obtained by oppression of the person who made it or as a result of anything said or done that was likely to render the confession unreliable. Oppression is defined to include torture, inhuman or degrading treatment, and the use or threat of violence, but there is no doubt that it includes other matters as well (such as excessively prolonged periods of questioning). This broad principle is supplemented by a much more detailed code of practice.

Prosecution

In countries whose legal system follows the English tradition, the function of prosecution is usually distinguished from that of investigation on the one hand and adjudication on the other. In most countries (although not in England until recently) the function of prosecution has been given to an official who is not part of either the police or the judicial system; a wide variety of terms are used to designate this official--district attorney in the state jurisdictions of the United States, procurator-fiscal in Scotland, and crown attorney in Canada are examples. The prosecutor may be an elected local official (as in the United States in most cases) or a member of an organization responsible to a minister of the national government. The first tasks of the prosecutor are to assess the information collected by the investigators, to determine whether there is sufficient evidence to justify the institution of criminal proceedings, and to decide whether there are any reasons why the public interest requires that a prosecution should not be undertaken.

In common-law systems the prosecutor is usually entrusted with extensive discretion in deciding whether to institute criminal proceedings. In part, this discretion arises out of the ambiguity of the criminal law; frequently a statute defining a particular criminal offense does not make absolutely clear what kind of behaviour it is intended to cover or includes a much wider range of circumstances than it was intended to prohibit. If this is so, the prosecutor must decide whether

the case he is dealing with falls within what was intended to be the scope of the law. Changing attitudes in the community toward particular kinds of behaviour may mean that a criminal prohibition, while remaining on the statute books, no longer reflects the sentiment of the community, and the prosecutor is no longer expected to bring charges against people who infringe it. In other cases, laws may be enacted without the usual exemptions from responsibility for those who commit the act unintentionally (offenses of strict liability). In such cases the prosecutor may nevertheless feel justified in not bringing proceedings against those who are technically guilty if they are in his view morally innocent.

The court system

Court systems and procedures reflect the history and culture of the country in which they have developed; there are many variations among different countries, or among different jurisdictions within the same country, regarding the way in which criminal cases are brought to trial.

CRIMINAL PROCEDURE IN ENGLISH-SPEAKING COUNTRIES

Each state of the United States has its own legal system, and, within the United Kingdom, England and Wales, Scotland, and Northern Ireland all have different arrangements for the conduct and procedure of criminal trials. These countries, however, generally follow what is called "adversarial" procedure, in which allegations are made by the prosecution, resisted by the defendant, and determined by an impartial trier of fact--judge or jury--who is generally required to find in favour of the defendant by acquitting him if there is any significant doubt as to his guilt. English criminal procedure, employing the adversarial method, is the model from which the court systems of many English-language systems have been developed (although Scotland evolved its own distinctively different rules independently); over the years the differences between, for instance, the English criminal courts and those of the typical U.S. state have widened in some aspects, but the same basic principles are still reflected in both countries. The court systems of most English-speaking countries provide two or more sets of criminal

procedure, to deal with the more serious and less serious cases, and a further set of procedures for hearing appeals against the decisions of courts of trial.

England.

All criminal cases brought to trial in England begin in the magistrates' court. The magistrates' court has a number of different functions to perform--to determine the mode of trial, to try the case if summary trial is chosen, and to deal with ancillary matters such as bail and the granting of legal aid. Although the expression "examining magistrates" is still found in the statutes, the magistrates have long ago lost any function in the investigation of the alleged crime; their function is now wholly concerned with the adjudicatory phase of the process. The police investigation is normally completed by the time the case comes before the magistrates' court for the first time. The magistrates themselves are for the most part laypeople chosen for their experience and knowledge of society. All are appointed by the central government on the advice of a committee (known as the Lord Lieutenant's Advisory Committee) for the particular county in which they are to sit. Magistrates, who are required to sit on an average of at least 14 days each year, develop considerable experience in their work, but they cannot be considered professionals. In large cities there are professional, legally qualified magistrates, known as stipendiary magistrates. The stipendiary magistrate can sit on his own, but lay magistrates may sit only as a bench of two or more. Lay magistrates are invariably attended by a legally qualified clerk to advise them on matters of law. The system of lay magistrates has existed in England and Wales since about 1360 and is generally an accepted part of the administration of justice.

The United States.

Criminal procedure in U.S. states follows a pattern derived from English traditions and principles, but with many variations. The lay magistrates play an insignificant role, if any, in the U.S. system, and the prosecutor (the district attorney) is a key courtroom figure. He determines the charges, which in turn may well determine whether the accused appears before a lower court (dealing with misdemeanours) or a higher court (dealing with felonies). The accused is offered bail in almost every

case, but he is not released unless he is able to deposit with the court either cash or security in the form of a bond, often posted on his behalf by a bondsman who charges a proportion of the amount of the bond. In some states it is common for an accused person to be released without bond on his own recognizance. The role of the examining magistrates in English criminal procedure may be played in the United States by the grand jury, whose task it is to examine the evidence produced by the prosecutor and, if warranted, to return an indictment. The deliberations and proceedings before the grand jury are normally conducted in private. When the case is brought before the trial court, it is often settled on the basis of a plea bargain made between the prosecutor and the defense lawyer, by which the accused pleads guilty to some of the charges and the prosecutor recommends a sentence that has been agreed upon beforehand. Plea bargaining, which can take many other forms, is more readily accepted in U.S. courts than in English courts as long as basic rules, designed to ensure fair dealing for the accused, are observed. If the case goes to trial before a jury, a major difference between the English and U.S. systems is seen in the procedure for the selection of the jurors. In a U.S. court the lawyers are allowed to question potential jurors about their beliefs and attitudes so as to exclude those who may be prejudiced. The selection of the jury in an important case may take almost as long as the hearing of the evidence. The rules of evidence are much the same as those followed in an English trial, with variations of detail, and the accused is normally represented by a lawyer paid for by the state if he cannot afford one himself. U.S. law allows a wider range of appeals, both within the state system and, if a question of constitutional rights is involved, by removal of the case to a federal court. It may be many years before the case is finally resolved beyond all dispute.

OTHER SYSTEMS OF CRIMINAL PROCEDURE

Continental Europe.

The jurisdictions of continental Europe follow methods of criminal procedure very different from those of the English-speaking world. Often described as the inquisitorial method, continental practice emphasizes the role of the judge, who is

normally responsible for calling and questioning all witnesses and who does not separate the process into two distinct phases of trial of guilt and sentencing. The tribunal may consist of several judges, or a combination of judges and lay assessors, who deliberate together on both conviction and sentence. The rules of evidence are generally less restrictive; materials that would be considered hearsay in common-law countries are often admitted, and information about the accused person's record is available to the tribunal. A major difference between the two traditions is that most European jurisdictions do not permit conviction on the basis of a plea of guilty; even though the accused is willing to admit his guilt, the court must investigate the evidence fully (although the admission is part of that evidence). A second major difference is that the decision of the tribunal is normally accompanied by a statement of reasons, which is never given for the verdict of a jury. (D.A.T.)

An institution without parallel in English law is the French unified magistracy, whose members are divided into assise ("seated," or the members of the bench) and parquet, or debout ("standing," or the prosecuting attorneys). It is a state prosecuting system in which the state acts as a party to the prosecution of civil and criminal cases.

Africa.

Prosecuting and sentencing systems in African countries in general follow those of the former colonial rulers from whom the legal systems are derived. In the common-law countries this means that, although there is everywhere state prosecution, considerable responsibility falls on the police forces to initiate prosecutions. Sentencing is the responsibility of the court that tries the case and convicts the defendant. In some countries, such as The Sudan and parts of Nigeria, where Indian legal influence was strong, versions of the Indian Criminal Procedure Code were adopted, in which the magistrate, rather than the police, takes charge of the investigation and levels charges.

Islamic countries.

Among Islamic countries of English and French colonial heritage, the modern states have adopted the procedure of the colonial countries that ruled them. Pakistan, for instance, which originally inherited the Indian Criminal Procedure Code, now has a procedural system very similar to that of England. It is an accusatorial system in which both sides present their oral arguments to an impartial judge. There is a competent and independent bar from whose ranks judges are chosen. This was amended in 1980 with the introduction of special Islamic courts and judges. On the other hand, Egypt's criminal system almost exactly mirrors that of France. The system is inquisitorial, and the judge has a much greater power to question and intervene and to determine the method of proceeding. There also exists the Niyaba, a system of state prosecutors very similar to the French parquet. Egyptian judges, unlike their English and Pakistani counterparts, are often career judges. In all categories of Islamic states there are not only ordinary criminal courts but often also police courts, which tend to deal with lesser criminal offenses, and military courts, which hear questions affecting security and military matters. In those states, such as Saudi Arabia and Iran, that claim to adhere totally, or almost so, to traditional Islamic law, Islamic judges, called qadis, exercise jurisdiction in Islamic courts. China.

The Chinese penal system broadly divides procedures and sanctions into criminal and administrative. "Crimes" are in China distinguished from "ordinary illegal acts." Crime is only that behaviour that is punishable by a court under the Criminal Law or other laws calling specifically for criminal punishment for violators. Ordinary illegal acts, however, can be punished administratively by nonjudicial bodies (such as the police) on their own initiative and according to their own less formal procedures. In general, administrative punishments cannot be appealed to a court. Milder sanctions, such as group crit icism, may be imposed by neighbourhood-level organs of local government. Thus, disruptions of social order can meet with a response by the state well before they reach a level called criminal. The concept of "circumstances" is of crucial importance to criminal procedure in China. Circumstances might, for example, mean the identity of the accused or the

victim, the existence of an official campaign against the particular type of crime involved, or such matters as whether a robber also beat his victim or whether the accused showed repentance. It is common in many countries for such factors to be taken into account in sentencing. Chinese law, however, differs in its provision for allowing circumstances also to bring an act within or entirely outside the coverage of the Criminal Law and, more importantly, the associated Criminal Procedure Law. It is only the latter, for example, that provides for a public trial by a court and the right to a defense. A person suspected of selling pornographic books may, if the police deem the circumstances clearly minor, be judged by the police and punished by up to 15 days' detention in a police station. In such cases there is no right to a defense and no appeal to a court, but the maximum punishment that can be imposed is much less than would be possible were the accused to be prosecuted under the Criminal Law.

Sentencing

In countries following the Anglo-American legal tradition, sentencing is a function that is distinguished from that of determining guilt or innocence and is normally the responsibility of the judge rather than of the jury, although in some parts of the United States the jury is empowered to determine the sentence. Most such systems of law traditionally give the judge a wide discretion in determining both the kind of penalty to be imposed (imprisonment, fine, probation) and its extent. As modern sentencing systems provide an increasingly wide range of forms of sentence, the choice of sentence becomes a more complex task. The extensive discretion involved in sentencing and the wide variety of different forms of sentence mean that in many cases there are complaints of disparity in the sentences passed on different offenders and of arbitrariness and idiosyncrasy in the decisions of individual judges; it is sometimes said that the sentence imposed on an offender may depend more on the judge before whom the offender appears than on the gravity of his offense or his record.

It has long been recognized that the quality of the decisions made by judges in sentencing depends on the information available to them. In a case in which there

has been a contested trial, the judge will have heard all of the evidence related to the immediate background of the offense but will not necessarily know much about the background of the offender. This gap is filled in many jurisdictions by a report (a presentence report or social-inquiry report) prepared by a probation officer and submitted to the court after the offender has been convicted or has pleaded guilty.

Parole

Origins of parole.

The word parole in French means "word," and its use in connection with the release of prisoners was derived from the idea that they were released on their word of honour. The practice of allowing prisoners to be released from prison before serving the sentence of imprisonment pronounced by the court goes back at least to 18th-century England. At that time almost all serious crimes (felonies) were punishable with death, but only a small proportion of those who were convicted of felonies were actually executed. The majority of those who were sentenced to death were pardoned by the king, but their pardon was granted on the condition that they consent to be transported to one of the colonies where labour was required--during the 17th and 18th centuries this was America and, following American independence, Australia. Eventually the courts were given power to pronounce sentences of transportation themselves, usually for a period specified in the sentence, but most sentences of transportation were modified by executive action. In particular, there developed the system of "ticket of leave," under which a convict detained under a sentence of transportation was allowed a measure of freedom, or the right to return to England, in return for good behaviour. When the sentence of transportation was abolished in the mid-19th century, the sentence that replaced it in English law, penal servitude, incorporated the same procedure under a different name, release on license. The prisoner sentenced to penal servitude could earn his release from the penitentiary, but not from the shadow of the sentence, by his good behaviour in custody. His release was decided by the executive government and was conditional on good behaviour outside prison; if another offense was committed, the prisoner could be returned to prison to serve

out the rest of the sentence (known as the remanet). In England the system of release from sentences of penal servitude became almost inflexible by the later years of the 19th century, with the result that all prisoners serving the sentence were released after serving a fixed and predetermined portion of it; in the United States at that time, however, the principle of the indeterminate sentence became widely accepted and eventually formed the basis of the sentencing laws of many states.

In those states where the indeterminate sentence was adopted, the law required a judge who decided to sentence an offender to a term of imprisonment to fix maximum and minimum limits of confinement; the actual date of his release, and the conditions, were then decided by an executive body usually known as the parole board, which had power also to revoke the offender's parole and return him to prison. The indeterminate sentence was seen to have a number of advantages over the more rigid form of sentence, in which the prisoner could work out his exact date of release from the moment he was sentenced. The indeterminate sentence allowed the authorities to observe the behaviour and attitudes of the offender while he was serving his sentence, and in particular the way in which these changed for the better; it provided an incentive to the prisoner to improve, in order to convince the authorities that he was ready for release. In addition to contributing in this way to the rehabilitation of the offender, the indeterminate sentence had a number of administrative advantages to the prison authorities. It provided a powerful sanction against misbehaviour--a prisoner who was violent or disruptive in prison knew that he risked losing the chance of release; it allowed the authorities to compensate for disparities in the sentences imposed by judges (often believed to be a source of friction and discontent among prisoners); and it provided a means by which the population of the prisons could be kept within limits.

Parole supervision.

An essential feature of parole is the supervision of the offender during the remaining part of the sentence after his release from prison. A prisoner who has been released on license is not free from all restrictions; he is normally required to

observe various conditions, which may be quite restrictive, dealing with such matters as where he lives and works or requiring him to undergo medical or psychiatric treatment. Failure to comply with these conditions can lead to the revocation of the parole, which means that the offender is returned to prison to serve out the remainder of the sentence e. Enforcement of theconditions, as well as the provision of help and counseling, is usually the responsibility of a probation or parole officer, to whom the paroled offender is required to report at stated intervals and who may have considerable power over the offender. (The license may include a condition requiring the offender to live at a place approved by the probation or parole officer, for instance, or require the officer's consent to a change of employment.) In many countries the decision of the supervising officer or the parole board to terminate the offender's license and require him to return to prison are not subject to appeal or judicial review, even though the consequences for the offender may be serious.

Theories and objectives of punishment

The proper objectives of a system of punishment administered by the official organs of the state have been the subject of debate among philosophers, lawyers, and legislators for centuries. A variety of different theories or objectives of punishment have been proposed, some differing only in minor degrees, some fundamentally in conflict with each other. The debate has often been confused by the fact that the same expression is frequently used to denote different ideas, which are not always clearly distinguished from one another; conversely, the same idea may be described by different names, resulting in increased confusion.

RETRIBUTION

In modern judicial systems the term retribution has acquired various shades of meaning. The key principle that all theories of retribution share is that there should be relation between the gravity of the crime and the severity of the punishment.

One theory of retribution proposes that punishment is not imposed in order to achieve a social objective (such as law-abiding behaviour in the future by the offender or others who witness his example) but is rather an end in itself. Those

who hold this view maintain that punishment does not require justification by intended results or effects. Retribution in this sense does not necessarily imply severity of punishment.

According to a second theory of retribution, punishment must be justifiable in relation to the gravity of the offense itself, even though there are other reasons (such as deterrence or treatment) that indicate a severer penalty than the gravity of the offense itself would warrant. Some commentators divide this limiting principle of retribution into two subdivisions. One is the principle that punishment should not be inflicted unless the recipient of the punishment has been found guilty of an offense, and the other is that the offender should not be punished more severely than his offense warrants. The first subdivision prohibits such practices as collective punishment imposed on whole communities or the taking of hostages from the general population, as has been practiced by occupying forces at various times. It also requires that the proper forms of the law be observed before punishment is inflicted. The second limb of the principle assumes that some scale can be drawn equating particular crimes with particular punishments. This is extremely difficult to do without resorting to a crude system of inflicting on the offender exactly the damage he has inflicted on the victim (which is in any event impossible in relation to many modern crimes, which have no specific victim). All that can be done is to draw up a conventional scale relating penalties to offenses; this may allow the severer penalties to be preserved for the graver offenses, but it will not in itself justify the relationship between any particular penalty and any particular offense, except by reference to the conventional values of the scale.

Retribution as a limiting principle can be distinguished from retribution as an educational principle, in that in the latter case enactment and implementation of the criminal law, and particularly the imposition of sentences, has the effect of providing a concrete example of society's values, which serves to reinforce those values among those who hold them and to instill them in those who do not. The member of the community who sees his moral values expressed in the judgment of the court in a particular case may feel more strongly committed to them than

previously; if he sees them ignored by the court, in the lenient treatment of an offender whose behaviour has violated the fundamental moral principles of society, he may come to question them himself and possibly feel less constrained by them. The principle assumes that a repeated failure of the courts to express such values would lead eventually to moral decline and the dissolution of society.

This aspect of retribution must be distinguished from yet another--the idea that the official organs of the state must punish offenders in order to satisfy the demand for punishment that is natural among members of the community, particularly among those who are the victims of the crime and who in the absence of official punishment administered by the state are likely to take the law into their own hands and seek revenge by direct violence. A variation on this interpretation of retribution is the concept of expiation--the idea that the offender should undergo punishment in his own interests, to discharge his guilt and to make himself acceptable to society again.

DETERRENCE

Like retribution, deterrence is a complex concept that informs several related theories of punishment. Criminologists distinguish between general and individual deterrence.

General deterrence.

A general deterrent is a punishment the object of which is to deter other persons from following the example of the offender, by fear of the same consequences that have been inflicted on him. The theory of general deterrence is not concerned with the future behaviour of the offender himself. It assumes that most crimes are rational and that potential offenders will calculate the risk of being caught, prosecuted, and sentenced in the same manner. Demonstrating the validity of the theory has proved difficult; general trends in crime and their relationship to particular sentencing policies (such as changes in the incidence of a particular crime after it has been treated more or less severely) are seldom a true indication of the effect of penalties as deterrents, as many other factors may be at work. Occasionally it is claimed that particular sentences have had a strong deterrent

effect, but usually closer inquiry reveals that this is open to doubt. One example of effective general deterrence is legislation designed to curtail driving after drinking alcohol; studies have suggested that mandatory penalties and a high probability of conviction of those detected have at least a temporary deterrent effect on a wide population. Individual deterrence.

Individual deterrence, unlike general deterrence, is aimed at the particular individual who is punished; the object is to teach him not to repeat the behaviour. This is the rationale of much informal punishment, such as is inflicted on children by parents as a part of their upbringing, as well as of formal sanctions administered by the authority of law. The effectiveness of this type of deterrence can be measured, at least in theory, by examining the conduct of the offender after the administration of the punishment to determine whether he has committed the offense again. Such studies can often be misleading, however, as in practice the only basis for determining that the offender has repeated the offense is a further conviction before a court. Because a high proportion of crimes do not result in convictions, many of those offenders who are not reconvicted after being punished in a particular way may have again committed offenses but avoided conviction.

A third sense of deterrence used by some writers coincides exactly with the idea of retribution as a form of moral education for the community as a whole, sometimes described as denunciation. Although this idea is closely associated with general deterrence through fear, and many sentences of the courts are intended to achieve both objectives simultaneously, there is an important difference between them. The idea of education through retribution, or denunciation, is aimed at the law-abiding person who is not tempted to commit crime; its object is to reinforce him in his rejection of lawbreaking behaviour of the kind in question. Most people do not steal because they know that stealing is dishonest and they consider themselves honest; a sentence on a thief reinforces them in this view. General deterrence through fear is aimed at those who do not necessarily reject the possibility of law-breaking behaviour on moral grounds but who do so on the basis of a careful calculation of the gains and risks involved. Those who are prepared to consider

stealing if they thought they could get away with it are frightened off by the example of the thief who is punished.

Deterrence shares with retribution the idea that punishments should be related in severity to the gravity of the crime. The principle of proportionality is central to the idea of deterrence, on practical grounds. If all punishments are the same, irrespective of the gravity of the crime, there may be no incentive to commit the lesser rather than the greater offense. The offender might just as well use violence against the victim of his theft, if the penalty for robbery is no severer than that for simple stealing.

INCAPACITATION

Incapacitation is an object of punishment that has been known since early times. The idea is simply that the offender should be dealt with in a manner that will make it impossible for him to repeat his offense--by execution or banishment in earlier times, in more modern times by execution or lengthy periods of incarceration. This is the only objective of punishment for which there is any certainty (although an offender incarcerated for this reason may escape or commit crimes within the prison). The difficulty arises in reconciling the idea with other principles, in particular those limiting retribution. In practice, it is limited to offenders who have committed crimes repeatedly (multiple recidivists) under what are known as habitual offender statutes (which permit courts to impose longer sentences on such offenders than are normally authorized for the offense) or to offenders who are designated as dangerous, in that there is reason to suppose them likely to commit grave crimes of violence in the future unless restrained. Because there is great difficulty in identifying with any certainty those offenders who are dangerous in this sense, the principle is controversial.

REHABILITATION

The most recently formulated theory of punishment is rehabilitation--the idea that through treatment and training the offender should be rendered capable of returning to society and functioning as a law-abiding member of the community. This idea began to establish itself in legal practice in the 19th century. Although

seen by many as a humane improvement on former practices, it did not always result in the offender's receiving a more lenient penalty than a retributive or deterrent philosophy would have given him. For many offenders, rehabilitation meant release on probation under some form of condition instead of a period in prison; for others it meant a longer period in custody undergoing treatment or training than would have been acceptable if it had been designated as punishment. One expression of the concept of rehabilitation was the indeterminate sentence popular in many U.S. jurisdictions, under which the length of detention was governed by the degree of reform exhibited.

Beginning in the 1970s, the concept of rehabilitation came under considerable criticism, and it is no longer as widely accepted as previously. The reasons for this growing skepticism are, first, the failure of criminologists to demonstrate that rehabilitation can be achieved in any systematic way. A second objection to rehabilitation as a theoretical basis for penal treatment is that sentences based on the concept typically give too much authority to the administrator, who may be empowered to decide to release or to continue to detain the offender, depending on his assessment of the offender's progress, which may itself be a vaguely defined measure. There have been cases in which this has led to gross abuse and the detention of offenders guilty only of minor crimes for long periods, out of all proportion to the gravity of their offenses, simply because of their inability or refusal to accept or adopt a subservient attitude to those in authority. A more fundamental objection to the concept of rehabilitation rests on a challenge to the criminological and political assumptions on which it is based. The idea that the offender can be successfully treated assumes that he is in some respect inadequate to comply with the legitimate demands of society on its citizens and is in need of additional training to supplement the deficiencies of his upbringing, education, or personality. Modern criminological theories that challenge this interpretation and portray criminal behaviour as a legitimate or at least predictable reaction to structural defects in society undermine the basis on which the concept of treatment is founded; this is reinforced by the similarities seen between the treatment of

delinquents and that of political dissidents in some countries. Some modern criminologists assert "the right to be different" and question the right of society to seek to change the values and ideas of individuals who choose to behave in a manner that brings them into conflict with the law.

THEORIES IN CONFLICT

It is obvious that, in the practical operation of a sentencing or penal system, these different theories often come into conflict. A lenient sentence (such as probation), designed to rehabilitate the offender, may fail to express society's rejection of his behaviour or to provide an effective deterrent to others. A sentence that requires the offender to submit to a compulsory program of treatment or training for a long period may in turn conflict with the idea of retribution as a limiting principle. A sentence of unusual severity, designed to make an example of the offender as a warning to others, is in conflict both with the principle of rehabilitation and with that of proportionality. A sentence whose object is incapacitation may fail to satisfy those who believe in rehabilitation or in proportionality. In practice the daily operation of any system of sentencing requires decision makers to choose between these different theories in different cases; no single theory provides a system suitable for all cases. The choice between them cannot yet be made on scientific grounds, and it may well be that criminology will never provide the information on which to base a scientific choice between the different objectives of punishment, if only because some of them rest on moral principles rather than on a supposed empirical effect.

PUNISHMENT IN OTHER SYSTEMS

Africa.

Sentencing courts in Africa stress the punitive and deterrent aspects of sentencing rather than the reformative. In consequence, sentences of imprisonment are often proportionately longer than is usual in Europe, for example, and fines are heavier. The legislation of African countries reflects the same approach to sentencing; capital punishment and in many cases corporal punishment are permitted and in some cases may be mandatory. A notable feature of the current legislation in

Tanzania is the imposition of minimum sentences for a variety of offenses, including dishonesty and theft of stock; the court in practice has no option but to impose a specified minimum prison sentence, which from 1963 to 1972 was to commence and end with 12 strokes of the cane. Islamic countries.

Traditional Islamic law divides crimes into two main categories. Five so-called hadd crimes are specifically mentioned, along with their appropriate penalties, in the Qur`an. All other crimes are called ta'zir crimes, and the punishment of these is left to the discretion of the qadi, although the books of Shari'ah law limit his discretion to certain traditional punishments. In Shari'ah there is found very little power to impose fines. The general punishments are imprisonment or corporal punishment. The traditional requirement of eyewitnesses to crimes considerably limited the application of the severest penalties. The imposition of fines is a recent innovation.

China. The chief goal of criminal punishment in China is reform. Secondary goals are specific deterrence (deterring the offender from repeating his crime) and general deterrence (deterring other would-be criminals).

An authoritative Chinese textbook on criminal law states that the goal of reform in criminal punishment is founded upon the historical mission of the proletariat to reform society and mankind. The thoughts of citizens are not their own affair; the government has the right and the duty to see to it that all members of society become "new men." The commission of a criminal act is, in a sense, evidence that the offender is in particular need of reform and hence justifies the use of particularly coercive measures. The notion that an offender incurs a debt to society that can be paid merely by serving a prison term is alien to Chinese penology. The state is keenly interested in changes in the offender's thinking during imprisonment. Thus, reform through labour and political study generally accompanies imprisonment for criminal offenses.

The primacy of reform over deterrence is intimately connected with Chinese theories on the causes of crime. Chinese criminology holds that crime can be reduced and eventually eliminated through thought reform, education, and the

perfection of socialist society. Criminal punishment is seen as merely a supplementary means to this end. According to this view, when the thought of all members of society has been reformed, there will be no more crime.

Treatment of juvenile offenders

Alternative Reformatory movement.

Early common law made no special provision for children who committed crimes. Provided that the child was over the minimum age for criminal responsibility--originally seven--and had "mischievous discretion"--the ability to tell right from wrong--the child was fully liable as an adult to the penalties provided by the law. During the 19th century, children who were liable criminally were imprisoned, and there are records of children being hanged as late as 1708. The need for special treatment of juvenile offenders was first recognized during the 19th century in the reformatory movement, the purpose of which was to establish institutions for young offenders, based on training, as an alternative to confinement in adult prisons. The idea of a special court system for juvenile offenders, the juvenile court, began to gain ground in the early years of the 20th century. Juvenile courts were established in England under legislation enacted in 1908, and most U.S. states had juvenile courts by the 1920s. Although there are significant differences between the concepts of the juvenile courts in the two countries, in both the juvenile court deals with criminal and noncriminal cases. The English juvenile court is essentially a magistrates' court, exercising the ordinary criminal jurisdiction of the magistrates' court over a limited age group of offenders--from 10 years (the minimum age of criminal responsibility) to 17. (Those under 14 are designated as "children"; those over 14 and under 17 are "young persons.") Offenders aged 17 and over appear in the normal adult courts, although special sentencing provisions apply to offenders under the age of 21.

The main difference between a juvenile court and an adult court in England is that the juvenile court has a much wider jurisdiction in terms of the offenses it can try. It can deal with a juvenile for any offense except homicide, although it is not bound to deal with a young person for a serious offense such as robbery or rape; on

such a charge he can be committed to the Crown Court for trial in the same manner as an adult. A child may be committed to the Crown Court for trial on a charge of murder or manslaughter. A juvenile may also be sent to an adult court--magistrates' court or Crown Court--for trial, if charged jointly with an adult (as, for example, in the case of a parent and child jointly charged with shoplifting). In such cases the adult court normally returns the juvenile to the juvenile court for sentencing. In addition to its criminal jurisdiction, the juvenile court may deal with children of any age up to 17 in what are called care proceedings--proceedings that are based on the idea that the child is in need of care protection, or control because one of a number of conditions is satisfied. These conditions include neglect or assault by parents but also include the fact that the juvenile has committed an offense (the "offense condition"). A juvenile who commits an offense can thus come before the juvenile court either in criminal proceedings or in care proceedings, although the court may not take action in care proceedings unless satisfied that the juvenile is in need of care, protection, or control; the fact that an offense has been committed is not in itself sufficient. This combination of two different roles in the juvenile court has been a source of difficulty and controversy for many years, particularly because the court in its criminal jurisdiction is required by law to "have regard to the welfare of the child or young person" and, if satisfied that it is necessary to do so, to remove him from unsatisfactory surroundings for his own good, irrespective of the gravity of his offense. A juvenile who appears before the juvenile court charged with a minor offense, if found to be in need of care or control on the basis of inquiries into his background, can therefore be removed from the care of his parents and possibly be required to reside in an institution (known as a community home), perhaps for a period of several years, and possibly under conditions of security. Proposals to change the image of the juvenile court so that its criminal aspect declined in importance and its welfare aspect was emphasized were enacted in 1969 but never fully implemented. The dual role of the juvenile court has thus been retained, but its authority has to some extent been transferred to an administrative body, the local authority. If the juvenile court makes a care order--in

practice the most powerful sanction it has available--its effect is to transfer to the local authority the parental rights over the child, and it is for the local authority to decide whether to allow the child to live at home with his parents (but subject to local authority control), to board the child with foster parents, or to require the child to live in a community home. The court has only a limited degree of control over this decision.

Sanctions.

The care order is only one of the sanctions available to the English juvenile court and is used only in a minority of the cases that come before it. Another measure, the supervision order, places the juvenile under the general supervision of a social worker but may require him to take part in a wide range of organized, constructive activities as intermediate treatment. A supervision order may also include restrictive requirements prohibiting the juvenile from certain activities or a curfew in the form of a "night restriction," requiring him to remain at home during the evening for a specified period. Juveniles may also be fined (although the court must usually order the parent to pay the fine) or be ordered to pay compensation, as in the case of an adult.

In U.S. jurisdictions, as in England, a high proportion of juvenile offenders are dealt with informally by means of cautions or counseling. Where a case is brought to court, the procedure is distinctively different from that of a criminal court. The court is not normally concerned with determining guilt or innocence so much as with making a finding of delinquency, which may be the basis for a disposition--either freedom in the community under supervision or confinement in a correctional facility for young people. In keeping with what was seen as the juvenile court's role as a welfare tribunal rather than a court of criminal jurisdiction, procedural standards in the United States were formerly rather elastic, but a series of decisions of the Supreme Court established that the basic rights granted by the Bill of Rights apply to proceedings in the juvenile court (although there is no right to jury trial). Most U.S. juvenile courts, like English courts, deal with cases of neglect as well as criminal cases and a further category of "status

offenses"--behaviour amounting to an offense only when committed by a juvenile (such as running away from home), which in England would fall within the scope of care proceedings.

Prisons

The idea of imprisonment as a form of punishment is relatively modern. Until the late 18th century, prisons were used primarily for the confinement of debtors who could not pay, of accused persons waiting to be tried, and of those convicted persons waiting for their sentences--death or transportation--to be put into effect. Although imprisonment was a sentence available to the courts in cases of misdemeanour, these were a small proportion of the cases tried, and the normal task of the assize courts in England was expressed in the fact that they were known as courts of "general gaol delivery"--delivery meaning the release of prisoners from the jails (to liberty or execution) rather than their commitment to the jails. The holding of accused persons awaiting trial remains an important function of prisons--in England about 20 percent of the prison population is unconvicted or unsentenced, while in some European countries (notably Italy) the proportion of the prison population that consists of unconvicted prisoners may be as high as 80 percent but since the late 18th century, with the decline of capital punishment, the prison has come to be used also as a place of punishment. The concept of the penitentiary was advocated in England during this period by Jeremy Bentham; at the same time in the United States penitentiaries were created first in Pennsylvania and then in New York. By the late 19th century the decline in the use of the death penalty and the abolition of transportation meant that a sentence of confinement (under a variety of names--in England it was known until 1948 as penal servitude) had become the principal sanction for most serious crimes.

DEVELOPMENT OF THE PENITENTIARY

The development of the penitentiary in the late 18th century was in part a reaction to the conditions of the jails of the period. Sanitation in English prisons at this time was such that disease was widespread among prisoners, who were generally held without any segregation according to sex or classification; outbreaks of "jail fever"

occasionally killed not only the prisoners but also the jailers, and even on occasions the judges and lawyers involved in their trials. At this time many prisoners in England were confined not in buildings but in the hulks of ships moored in the Thames River and elsewhere; in theory they were waiting to sail for Australia under a sentence of transportation, but in practice many of them served their sentences in the hulks and were released without ever leaving the country. The appalling conditions in the many local prisons of late 18th-century England were exposed by John Howard, who had traveled throughout Britain, and later throughout Europe, for the preparation of his book The State of the Prisons in England and Wales (1777). Reaction against the gross severity of the former penal system of death and transportation and against the physical conditions of the jails led in England to the building of "convict prisons" by the central government; the local jails remained under the control of local authorities until 1877. In that year the whole prison system of England and Wales was brought under central government control, to be administered by a body known as the Prison Commission.

In the United States the prison system is more complex; offenders who are sentenced by federal courts for crimes against the federal criminal code serve their sentences in federal penitentiaries managed by the federal government, but the majority of offenders who are in custody are in state or local institutions, which form part of the penal system of the particular state. This consists of one or more state penitentiaries, possibly supplemented by a number of institutions offering a lower degree of security, such as prison camps or farms, and local jails, each managed and financed by the local jurisdiction in which it is situated. The principal function of the jail is still that of holding persons awaiting trial, but usually short sentences (less than 12 months) are served in the local jail rather than the state penitentiary.

THE PRISON POPULATION

Growth trends.

Concern over prison conditions has not diminished over the years. Problems of security and the protection of prisoners from violence on the part of other prisoners have been compounded by the difficulties arising from overcrowding, as prison populations in most countries continue to grow. Increasing prison populations have been a common feature of most industrialized societies in the era since World War II. In England in 1880 the prison population stood at 32,000; as the prisons came under central government control, there began a long period of decline, probably the result of changes in sentencing laws and practices. By the end of World War I the daily average prison population had declined to about 10,000. It remained stable during the interwar years, rising and falling slightly from one year to the next, but after World War II there began a period of steady increase that has continued unabated. From a daily average figure of about 12,000 in 1945, the prison population grew relentlessly each year, despite a variety of changes in the law designed to contain it. By the 1960s it had reached 30,000--the level of the 1880s--and by 1976, despite the introduction of a parole system and suspension of sentences of imprisonment, the figure of 40,000 was exceeded. Since then it has never dropped below that figure, and by 1984 the population touched 45,000 on occasions. A similar situation has occurred in the United States. A total of 250,000 persons were incarcerated in 1975; by the end of 1980 this had grown to 315,000.

The rise in prison populations is attributed to a variety of factors, but probably the most significant is the rise in the incidence of reported crime and in the number of offenders brought before the courts. In England, both reported crime and the number of persons convicted have risen faster than the prison population. One country that has followed a different trend in the prison population is The Netherlands, where the prison population was halved from 1950 to 1975. This reduction has been attributed by criminologists primarily to a shortening of sentences passed by courts in The Netherlands for common offenses.

The people who make up the populations of most prison systems have many characteristics in common. They are predominantly male--in England males outnumber females by 28 to 1 (although the number of women in prison is rising at

a higher rate than the number of men)--and relatively young--nearly 70 percent of those in custody are under the age of 30. Most offenders in prison have a number of previous convictions; the offenses they have committed are most commonly burglary, theft, violence, or robbery. A similar picture is revealed by U.S. statistics; the most common offenses for which prisoners are in custody are burglary and robbery.

Types of institutions.

Prisoners are distributed among a variety of types of institution. In the United States most prisoners serving longer sentences are held in state prisons, which are usually large maximum-security buildings holding more than 1,000 offenders in conditions of strict security. Young offenders usually are detained in separate institutions, often designated under names that imply that their purpose is treatment or correction rather than punishment. Women are normally held in separate institutions. Prisoners who are not considered a danger to the community may be confined in low-security or open prisons.

In England, as in the United States, most prisoners are held in prisons constructed more than a century ago. Prisons are classified administratively as local or central prisons. Local prisons serve a variety of purposes--holding prisoners awaiting trial or sentencing and prisoners serving shorter sentences (up to about 18 months). There the worst overcrowding occurs. Prisoners serving longer sentences are detained in central prisons, dealing exclusively with similar cases. For security, prisoners are classified into four categories, from A (prisoners likely to attempt escape, and constituting, if successful, a significant danger to the public) to D (prisoners who can be trusted to work in conditions of minimal security). Central prisons cover a range from maximum-security institutions to medium-security prisons, where the degree of security is less intense; and to open prisons, where physical security is minimal and there is normally no obstacle to a prisoner's absconding. In some European countries a further category

of institution is available to accommodate prisoners who are allowed to serve their sentences intermittently, usually over a series of weekends. Younger offenders in

England (in the age group 15-21) were until recently held in Borstal institutions, named after the village in Kent where the first one was operated. For many years these institutions were admired as an example of practical rehabilitation through training, but declining enthusiasm for this concept, and disillusionment with its effectiveness, led to its replacement with that of "youth custody." Another feature of the English prison system is the detention centre. These institutions for young males serving sentences that must not exceed four months are based on the principle of vigorous discipline and physical activity, popularly known as the "short sharp shock"; research has failed to show, however, that it is an effective deterrent to further crime.

Prisons have been described as total institutions, in which every aspect of life is subject to control. In addition to daily routines such as mealtimes, times of rising and retiring, and bathing, many other aspects of the prisoner's life are subject to control. In part this control forms the deprivation of freedom that is the essence of imprisonment, and in part it is a necessary adjunct as a means of maintaining security, controlling the introduction of weapons or contraband substances, and preventing escapes. Most prisons limit the number of visits that a prisoner may receive from his family or friends. In England the Prison Rules allow a convicted prisoner one visit every four weeks, although the prison governor may increase or limit visits at his discretion. Only relatives and friends of the prisoner may visit him, although adequate facilities must be available for visits by legal advisers if the prisoner is engaged in any litigation (for instance, divorce proceedings). Visits normally take place within the sight of an officer, and in some cases within his hearing. In many prisons, visits are conducted with the prisoner sitting on one side of a table and his visitor on the other, with a wire mesh partition between them; the visitor may be searched for contraband. In other prisons the conditions for visiting may be less restrictive--the visitor and the prisoner may be allowed to meet in a room without any physical barrier but still in the sight of officers. Conjugal visits (in which the prisoner's spouse comes to stay with the prisoner for a period of several days) are not permitted in England, but some U.S. states do permit them.

Correspondence of prisoners in England is subject to censorship by the prison authorities, and prisoners may not write more than one letter each week.

Control of the prison is maintained by a number of disciplinary sanctions, which may include forfeiture of privileges, confinement within a punishment block or cell, or the loss of remission or good time (time deducted from the sentence as a reward for good behaviour). The procedures for the imposition of sanctions on prisoners have been improved in both England and the United States, in part as a result of actions taken through the courts. Generally, prisons are governed by rules setting out a code of conduct and listing prohibited behaviour; the code must be given to the prisoner on his arrival in the prison. Typically, the prohibited offenses include mutiny and violence to officers; escaping, or being absent from a place where the prisoner is required to be; and possessing unauthorized articles. The rules may also include one or more generally defined offenses (such as the English "offence against good order and discipline") that leave much scope for interpretation. Disciplinary sanction may be imposed by the prison administrator or governor in minor cases, but the imposition of a more serious sanction--e.g., loss of remission or good time--requires a more formal disciplinary hearing before a committee or board, which will follow the basic rules of procedure in a court of law.

Prisoners' rights.

The idea that the prisoner has rights that may be protected by actions in the courts has been developed particularly in the United States, where actions brought under the provisions of the U.S. Constitution (notably the Eighth and Fourteenth amendments) have established that as a general principle prisoners are entitled to the protection of the Constitution and that interference with rights guaranteed by the Constitution to citizens in general requires special justification. In some cases, courts have ordered state prison administrators to make major improvements in prison conditions or to close down particular institutions, but not all of these decisions have been effectively enforced. In England, in the absence of a written constitution, prisoners resorting to the courts have relied on the general principles

of administrative law, which require fair procedures by disciplinary bodies. Although many actions brought by prisoners have been unsuccessful, prison disciplinary procedures have been improved as a result of such litigation.

For many prisoners the worst pressures arise not from the prison authorities but from fellow prisoners, particularly in overcrowded institutions where prisoners are forced to share cells and supervision is limited. Some criminologists have claimed that there exists a prison subculture, standing opposed to the official hierarchy of the prisons, which demands the loyalty of the prisoner and expects him to conform to a series of informal rules, enforcing his compliance by violence and social pressures. Certain types of prisoner are particularly likely to be treated with violence by other prisoners--those who have committed sexual offenses against children (known in English prison argot as "nonces") or law enforcement officers sentenced for corruption or similar matters. In the English prison system such offenders may be allowed to go into solitary confinement for their own protection. Racial conflict is a major problem in many U.S. prisons, and riots have occurred in prisons in the United States and other countries, usually as a result of grievances over the management of the prison, disparity of sentencing, and the uncertainties of the parole system.

One innovation of some importance in England is the creation of an Inspectorate of Prisons. The chief inspector of prisons and his assistants have the responsibility of inspecting all prisons in England and of reporting directly to the home secretary (rather than to the prison department of the Home Office). An annual report is published each year by the Inspectorate, dealing with the general problems of the system, and individual reports, some of them very critical, have been published on the conditions and management of particular institutions.

The death penalty

IN ENGLISH LAW

Death was formerly the penalty for all felonies in English law. In practice the death penalty was never applied as widely as the law provided, as a variety of procedures were adopted to mitigate the harshness of the law. Many offenders who committed

capital crimes were pardoned, usually on condition that they agreed to be transported (or to transport themselves) to what were then the American colonies; others were allowed what was known as benefit of clergy. The origin of benefit of clergy was that offenders who were ordained priests (clerks in Holy Orders) were subject to trial by the church courts rather than the secular courts; if the offender convicted of a felony could show that he had been ordained, he was allowed to go free, subject to the possibility of being punished by the ecclesiastical courts. In medieval times the only proof of ordination was literacy, and it became the custom by the 17th century to allow anyone convicted of a felony to escape the death sentence by giving proof of literacy. All that was required was the ability to read (or recite) one particular verse from Psalm 51 of the Bible, known as the "neck verse" (for its ability to save one's neck); most offenders learned the words by heart.

In 18th-century England concern with rising crime led to many statutes either extending the number of offenses punishable with death or doing away with benefit of clergy for existing felonies, which as a result became capital. By the end of the 18th century English criminal law contained about 200 capital offenses. The application of the death penalty was extremely erratic, as in any capital case the judge was entitled to reprieve the offender so that he could petition for mercy; but the judge was not obliged to do this, and if he decided to "leave the offender for execution," the death sentence was normally carried out immediately after the closing of the assize, without appeal. In practice, many offenders who were convicted of capital crimes escaped the gallows as a result of reprieves and royal pardons, usually on condition of transportation, and many others who were charged with capital crimes were acquitted against the evidence, because the jury was unwilling to see the death penalty applied in a minor case.

The erratic application of the death penalty in the late 18th and early 19th centuries led to demands for reform, both from humanitarian reformers, such as Sir Samuel Romilly, and from those who were more concerned with the effectiveness of the legal system and who could see that the very severity and arbitrariness of the law

and its administration undermined its deterrent effect. Between 1820 and 1840 most of the capital statutes were repealed, and by 1861 only four offenses retained the death penalty--murder, treason, arson in a royal dockyard, and piracy with violence.

Until the mid-19th century executions in England were public, and throughout the 18th century great crowds attended the regular executions in London and other cities. Often an execution was followed by scenes of violence and disorder in the crowd, and it was commonly believed that pickpockets were busy among the spectators at executions. Public opinion eventually turned against the idea of executions as spectacles, and after 1868 executions were carried out in private in prisons.

Although treason remained a capital crime in England, and persons convicted of treason were executed after both world wars, in practice the only capital crime for which criminals remained liable to be executed was murder. (Arson in a royal dockyard ceased to be capital in 1971.) From the 1930s until the mid-1960s, reformers campaigned for the abolition of the death penalty for murder. One attempt came close to succeeding in 1947, and later the government appointed a royal commission to consider the question; the report of this commission, the Royal Commission on Capital Punishment 1953, remains one of the most significant and comprehensive accounts of the question. Following its publication, a number of controversial executions, and a further parliamentary attempt to abolish the death penalty altogether, Parliament enacted in 1957 a statute restricting the death penalty to certain types of murder, known as "capital murders"--murder in the course of theft, murder of a police or prison officer in the execution of his duty, murder by shooting or causing an explosion, and murder on a second occasion. All other murders were to be punished by a mandatory life sentence (although murderers sentenced to life imprisonment were eligible to be released on license at any time during their sentence).

The operation of the system of capital murder created great dissatisfaction, as it led to some executions that the public viewed as unjustified, while other types of

murderers escaped the death penalty simply because of the method used to commit the crime (in particular, deliberate poisoners were not subject to the death penalty, but the emotional murderer who had happened to seize a gun was liable to execution). Another objection was the fact that the liability to the death penalty might depend on a narrow question of law--such as whether a murder committed by a burglar escaping from the scene of the burglary occurred "in the course or furtherance of theft" if the theft was already complete. These objections led to a further move for change, and in 1965 the Murder (Abolition of Death Penalty) Act was passed, abolishing the death penalty for all murders and replacing it with a mandatory life sentence in all cases. The judge was given the power to recommend that the offender sentenced to life imprisonment should not be released before he had served a certain minimum period. Despite two parliamentary motions to restore the death penalty, both of which failed, this remained the position in England and Wales and Scotland in the mid-1980s. Northern Ireland retained the death penalty under a law designating certain murders as capital but abolished it in 1973.

IN THE UNITED STATES

In the United States, where the existence of the death penalty is primarily a matter of state law, capital punishment was never as widely provided as in 18th-century England, but it was permitted by many states for murder and in some states for offenses such as rape and kidnapping. Executions were common; between 150 and 200 persons were executed each year in the decade before World War II. In the postwar years the number of executions declined to about 50 each year by the late 1950s. During the 1960s doubts grew as to whether the application of the death penalty was constitutional; the question was raised as to whether execution was "cruel and unusual punishment" of a kind forbidden by the Eighth Amendment to the Constitution or whether it violated the requirement of the Fifth and Fourteenth amendments that all persons within the United States should be afforded equal protection under the law. These doubts led to a complete cessation of executions for a decade, until the constitutional issues were settled by the Supreme Court of

the United States in 1972 in the case of Furman v. Georgia, although this turned out to be a confusing ruling. The Supreme Court ruled that the death penalty itself did not violate the Constitution but that the manner of its application in many states did. It was shown that capital punishment was likely to be imposed in a discriminatory way and in particular that blacks were far more likely to be executed than whites. The decision in Furman v. Georgia had left uncertain the precise requirements of the Constitution for a valid death penalty statute, except that it required a system for applying the death penalty that would not be discriminatory against any racial or other minority.

Some states enacted legislation making the death penalty mandatory in all cases of convictions for the crime in question, on the assumption that, if there was no discretion in the application of the penalty, there could be no question of discrimination in its application. Other states enacted statutes that provided for the death penalty to be imposed only after a special hearing, at which matters of mitigation and aggravation were to be considered, so that the discretion would be exercised in a systematic rather than an arbitrary manner. The constitutionality of these new statutes was considered by the Supreme Court in a series of decisions in 1976, which decided that laws making the application of the death penalty automatic were unconstitutional but that those providing a framework for the exercise of discretion in a structured manner were constitutional. The decision upheld the death penalty statutes of some states; in the light of the decision, other states enacted new legislation providing for the application of the death penalty in a manner indicated by the Supreme Court. About two-thirds of the states now have provisions in their laws for the death penalty for murder; although some states provide the death penalty for other crimes, there is doubt about its constitutionality in these cases.

One effect of the doubts over the constitutionality of the death penalty, together with the length of time needed to exhaust the appeal procedures available in the United States, was that the informal moratorium on the carrying out of death sentences that had begun in 1967 came to an end. The first execution under the

new legislation took place in 1977, but many who were sentenced to death after 1976 contested the validity of the convictions or sentences, with the result that a large number of convicted offenders were held on "death row" in U.S. prisons, waiting in some cases for years to know if they were to be executed.

IN OTHER COUNTRIES

Continental Europe.

Most European countries have abolished the death penalty for murder; it remains on the statute book for peacetime offenses only in Belgium but is never carried out in practice. (D.A.T.)

Africa.

The death penalty has been generally retained by African countries, and (in default of an active public opinion) there are no movements to abandon it. Indeed, African governments increasingly resort to the death penalty as the sanction for new classes of offenses, such as armed robbery. Military takeovers often lead to the execution by firing squad of those found guilty of serious "crimes" committed before the takeover.

China.

Despite the penological ideal of reform, the death penalty is widely used in China. Most executions appear to be for murder, rape, and robbery with violence. In 1981 the number of offenses carrying a possible death penalty was increased to include theft, bribery, embezzlement, molesting women, gang fighting, drug trafficking, pimping, and teaching criminal methods. The Criminal Procedure Law adopted in 1979 originally provided for all death sentences to be approved by the Supreme People's Court, China's highest judicial organ, but in 1981 this general requirement was removed in cases of murder, rape, robbery, and a number of other crimes, such as breaching dikes, that involved danger to the public.

Death sentences in China can be for immediate execution or for suspended execution, whereby the condemned is given a two-year reprieve. If a person shows evidence of reform and repentance, the sentence may be commuted at the end of the two years to a life sentence or a fixed term of imprisonment.

A distinctive feature of the death penalty in China has been the use of "mass sentencing rallies" to publicize exemplary cases. Condemned prisoners have frequently been paraded in public before their execution. The Criminal Procedure Law provides that the execution itself not be public, but this rule has not been universally observed.

THE CONTINUING CONTROVERSY

Arguments for and against the death penalty take many forms. Those in favour of its retention or reintroduction for murder claim that it has a uniquely potent deterrent effect on potentially violent offenders for whom the threat of imprisonment is not a sufficient restraint; that death is the only penalty that adequately reflects the gravity of murder; that prolonged detention over decades is a harsher penalty than death; that execution is the only sure means of preventing a murderer from being released or escaping and committing more murders. They may argue also that to keep murderers in prison for long periods is uneconomical. Those against the death penalty point out that there is no evidence of its being a more potent deterrent than the threat of a sentence of life imprisonment; that the death penalty tends to be imposed in a discriminatory manner on the poor and on members of minority groups; that it creates the risk that an innocent person may be executed; that it prevents any possible rehabilitation of the offender; that it lowers the community and the state to the same level of behaviour as the criminal; that violence used by the state in the form of capital punishment breeds violence by criminals and brutalizes those who have to administer it; that the death penalty distorts the administration of the criminal law and sensationalizes trials. Criminologists have never succeeded in producing convincing evidence to resolve these issues. Many of the arguments involved in the debate are questions of morality and personal conviction that are not within the scope of criminological research.

Alternatives to prison

In most criminal justice systems the majority of offenders are dealt with by means other than custody--by fines and other financial penalties, by probation or

supervision, or by orders to make reparation in some practical form to the community.

Fines.

The most common penalty is the fine. In 1984 in England, for instance, 42 percent of all criminal offenders (excluding motoring offenders) were fined, the same percentage were dealt with by various other means not involving custody, and 16 percent were imprisoned in one manner or another. The fine is a simple penalty that avoids the disadvantages of many other forms of sentence; it is inexpensive to administer and does not normally have the side effects, such as social stigma and loss of job, that may follow imprisonment. Fining has limitations, however. There are dangers that the imposition of financial penalties may result in more affluent offenders' receiving penalties that they can easily discharge, while less affluent offenders are placed under burdens that they cannot sustain. In some cases, it has been suggested, the more affluent offender, who is able to pay a very large fine, may be able to persuade the court to fine him in circumstances where any other offender would be sent to prison; such discrimination is likely to lessen respect for the legal system. Other problems arise when courts have to deal with offenders who have no financial resources or with those whose incomes are too small to allow them to pay anything more than a derisory fine. Some countries, notably Sweden, solve this problem by allowing the court to calculate the fine not in terms of a sum of money but as a number of days earnings.

The problem of lack of means is to some extent related to that of the enforcement of fines; a significant number of offenders who are fined have to be brought back to court for nonpayment of the fines imposed on them. If the court is satisfied that the offender has failed to pay as a result of willful neglect or culpable default, and that other means of securing payment are unlikely to succeed, he may be committed to prison. The other means include seizure and sale of the offender's property (distress) or seizure of any funds he may have in a bank or savings account (garnishee order). The length of time for which an offender may be committed to prison for deliberate nonpayment of a fine depends on the amount

outstanding. Each year about 20,000 offenders are imprisoned in England for deliberate nonpayment of fines (this represents about 1 percent of those who are fined), but most of these are imprisoned for very short periods--one or two weeks is typical. If the offender is able to pay the amount outstanding, he is entitled to immediate release; if he pays a part, the term of imprisonment is reduced proportionately. Restitution.

Related to the fine is an order to pay restitution (in some countries termed compensation). The principle of restitution is popular in some countries as an alternative to punitive sentencing, but there are some drawbacks. One is the possibility, as in the case of the fine, that the more affluent offender may receive favourable treatment from the court because he is able to pay restitution (particularly if he pleads that he should not be sent to prison in order to allow him to continue to earn the money with which to pay restitution to the victim). The second drawback is that such schemes do not help all victims of crime. Only those who are the victims of crimes for which the offender is caught and convicted and has the funds to pay restitution are likely to be recompensed. Even when an offender is ordered to pay restitution, it is often by installments over a long period. Victims of crimes of violence in some countries--such as England and Canada--are entitled to restitution from public funds, whether or not the offender is detected or has the resources necessary to compensate him. The English scheme is administered by the Criminal Injuries Compensation Board, to whom the victim applies, showing evidence of the violence against him and the extent of the loss he has suffered as a result of the crime. If the board is satisfied that the crime has occurred and that the claim is reasonable, the victim is compensated out of public funds in a single payment, sometimes in a large amount. The scheme has a number of limitations--the relatives of murdered people do not normally receive anything, unless they were financially dependent on the victim or the victim was a child under 17 (when a token amount is payable); nothing is paid at all if the total extent of the injury is such that the amount which would be awarded is less than 400; and

nothing is payable at all if the victim in any way provoked the crime or has himself a record of criminal offenses.

Other penalties.

There are many ways of dealing with offenders that do not involve the payment of money. One is probation, a system that takes many different forms in different jurisdictions but that essentially involves the suspension of sentence on the offender subject to the condition that he is supervised while living in the community by a probation officer and possibly agrees to comply with such other requirements as the court may think appropriate. Usually, if the offender complies with the probation order and commits no further offense while it is in force, no other penalty is imposed, but if he breaks the requirement of the order or commits another offense, he can be brought back before the court and punished for the original offense as well as the later one. In many U.S. states probation is combined with a suspended sentence, so that the sentence the offender will have to serve if he breaks the order is fixed in advance; in England the sentence is not fixed in advance, and the court has complete discretion if there is a breach to sentence the offender for the original crime in light of his later behaviour. English law also allows suspended sentence of imprisonment for a specified period (not more than two years), on condition that the offender commit no further offense during the period of suspension. This is different from a probation order, as no supervision is required and no other conditions may be included in the order.

Offenders who are found to be suffering from mental illness may be committed to a mental hospital rather than a penal institution. English law, for instance, allows a criminal court to make a hospital order against an offender whose mental condition warrants his detention for treatment, provided that there is medical evidence to that effect before the court. Alternatively, the court may make a probation order with a condition that the offender undertake psychiatric treatment. An offender who is detained in a hospital may apply to the Mental Health Review tribunal, which can order his release if it considers that his detention is no longer justified.

The concept of reparation has gained in popularity in a number of jurisdictions. Under this method, the offender makes good the damage he has done through his crime, not by paying money but by providing services to the victim directly or indirectly through the community. In England this takes the form of the community service order, under which the court is empowered to order anyone who is convicted of an offense that could be punished with imprisonment to perform up to 240 hours of unpaid work for the community, usually over a period of not more than 12 months. The consent of the offender is necessary before the court can make such an order, to avoid allegations that it amounts to forced labour. Typically, the offender carries out work in his leisure time, under the direction of the probation service. The kind of work involved varies according to the area, the time of year, and the abilities of the offender; in some cases it may involve heavy physical labour, but in others it may require such work as the provision of help to handicapped people. If the offender completes the hours of work ordered by the court, he receives no further penalty, but if he fails to carry out the work, without reasonable excuse, he can be resentenced for the original offense. Although follow-up studies of offenders given community service orders have not shown that this method is more effective than other forms of sentence in preventing further offenses (the same proportion of offenders who receive community service orders are convicted of later offenses as are those sentenced in other ways), the community service order is widely judged to be a successful innovation, and several other countries have adopted systems based on the same principle. It is less expensive to administer than imprisonment, less damaging to the offender and his family, and more useful to the community, and the majority of offenders complete the order satisfactorily, whatever their subsequent behaviour may be. There are some doubts about the extent to which the availability of community service as an alternative to prison weakens the deterrent effect of the criminal law, but there can be no doubt that community service has become an established sentencing alternative.

Other alternatives to prison are based on the idea of preventing an offender from committing further offenses, without necessarily confining him in a prison. The most familiar power of this kind is that of disqualifying an offender from driving a motor vehicle or from holding a driver's license. This power is available under the laws of most countries to deal with those offenders who either commit serious driving offenses, such as driving while intoxicated, or who commit repeated but less serious offenses, such as speeding. In many countries there exists a system in which the offender is awarded a number of points each time he commits a motoring offense; when the number of points accumulated reaches a certain figure, he is automatically disqualified for a specified period. Some countries allow courts to disqualify from driving those offenders who have used motor vehicles in commission of the crime for which they are being sentenced, with the aim of hindering the offender from committing further such offenses. Although attractive in the abstract, this seldom works well in practice, as the absence of a driver's license may well prevent an offender from finding work after release from prison; as a result he may be likely to commit further crimes. Other forms of disqualification may be imposed on offenders convicted of particular types of crimes: a fraudulent company director may be disqualified from being involved in the direction of a company, a corrupt politician may be disqualified from holding public office, or a parent who sexually abuses his children may be deprived of parental authority over them.

Crime and social policy

Increasing crime appears to be a feature of all modern industrialized societies, and no developments in either law or penology can be shown to have had a significant impact on the problem. The effect of crime on the quality of life cannot be measured simply in terms of the actual incidence of crime, because the fear of crime affects far more people than are likely to become actual victims and forces them to accept limitations on their freedom of action. Paradoxically, many social changes that are perceived as progress may lead to further escalation in the incidence of crime--economic progress, producing greater wealth, almost always

leads to greater opportunities for crime in the form of more goods to steal or enhanced possibilities for successful fraud--and an increase in individual liberty may have similar effects, as the older constraints on behaviour are discarded. Crime is least likely to be a serious problem in a society that is economically undeveloped and subject to restrictive religious or similar restraints on behaviour. For modern urbanized society, in which economic growth and personal success are dominant values, there is little reason to suppose that crime rates will not continue to increase.

Judicial and Arbitrational Systems

Introduction

This article deals primarily with the operations of the judicial branch of government. It explores some of the fundamental relationships of this branch with legislative and executive branches and analyzes the functions, the structure and organization, and, finally, the key personnel of courts, such as judges and juries. This article also treats arbitration, another legal means of resolving disputes.

The approach is comparative, contrasting and comparing the systems of the two predominant legal traditions of the contemporary world: first, that of the common law, represented by England, the United States, Canada, Australia, and other nations deriving their legal systems from the English model; and, second, that of the civil law, as represented by nations of western Europe and Latin America and certain Asian and African nations that have modelled their legal systems on western European patterns. Reference is made to the legal institutions in the former Soviet Union and in eastern European nations. A separate section deals more specifically with judicial systems in Communist countries.

Functions of courts

KEEPING PEACE

The primary function of any court system in any nation--to help keep domestic peace--is so obvious that it is rarely considered or mentioned. If there were no agency to decide impartially and authoritatively whether a person had committed a crime and, if so, what should be done with him, other persons offended by his conduct would take the law into their own hands and proceed to punish him according to their uncontrolled discretion. If there were no agency empowered to decide private disputes impartially and authoritatively, self-help, quickly degenerating into physical violence, would prevail and anarchy would result. Not even a primitive society could survive under such conditions. All social order would be destroyed. In this most basic sense, courts constitute an essential element in society's machinery for keeping peace.

DECIDING CONTROVERSIES

In the course of helping to keep the peace, courts are called upon to decide controversies. If, in a criminal case, the defendant denies committing the acts charged against him, the court must choose between his version of the facts and the prosecution's; and if he asserts that his conduct did not constitute a crime, the court must decide whether his view of the law or the prosecution's is correct. In a civil case, if the defendant disputes the plaintiff's account of what happened between them--for example, whether they entered into a certain agreement--or if he disputes the plaintiff's view of the legal significance of whatever occurred--for example, whether the agreement was legally binding--the court again must choose between the contentions of the parties. The issues presented to, and decided by, the court may be either factual, legal, or both.

It would be a mistake, however, to assume that courts spend all of their time deciding controversies. Many cases brought before them are not contested. They represent potential, rather than actual, controversies in which the court's role is more administrative than adjudicatory. The mere existence of a court renders unnecessary any very frequent exercise of its powers. The fact that it operates by known rules and with reasonably predictable results leads those who might otherwise engage in controversy to compose their differences.

Most people arrested and charged with crime in the common-law world plead guilty. If they do so understandingly and without coercion of any sort, there is no need to determine guilt, for the sole question is whether the defendant should go to jail, pay a fine, or be subjected to other corrective treatment. In civil-law countries some judicial inquiry into the question of guilt or innocence is required even after a confession. But the inquiry is brief and tends to be perfunctory. The main problem to be resolved, usually without contest, is what sentence should be imposed.

The vast majority of civil cases are also uncontested or, at least, are settled before trial. The court keeps the calendar moving, sometimes encouraging settlement, and decides such questions of law or fact as are presented by the parties; but the number of cases actually tried is small compared to the number settled.

Most divorce cases are uncontested, both parties usually being anxious to terminate the marriage and often agreeing on related questions concerning support and the custody of children. All the court does in such cases is to review what the parties have agreed upon and give its official approval.

Many other uncontested matters come before courts, such as the adoption of children, the distribution of assets in trusts and estates, and the setting up of corporations. Occasionally questions of law or fact arise that have to be decided by the court, but normally all that is required is judicial supervision and approval.

JUDICIAL LAWMAKING

As courts decide controversies they create an important by-product beyond the peaceful settlement of disputes, that is, the development of rules for future cases. Law is thus made not only by legislatures but also by the courts.

To an extent that varies greatly between common-law and civil-law nations, all courts apply preexisting rules formulated by legislative bodies. In the course of doing so, they interpret those rules, sometimes distorting them, sometimes transforming them from generalities to specifics, sometimes filling gaps to cover situations never considered by the original lawmakers. The judicial decisions embodying these interpretations then become controlling for future cases, sometimes to the extent of virtually supplanting the legislative enactments themselves.

This is one aspect of the doctrine of precedent, or, as it is sometimes called, stare decisis (literally, "to stand by decided matters"). Judges follow earlier decisions, not only to save themselves the effort of working out fresh solutions for the same problems each time they recur but also, and primarily, because their goal is to render uniform and stable justice. If one individual is dealt with in a certain way today, the theory is that another individual engaging in substantially identical conduct under substantially identical conditions tomorrow or a month or year hence should be dealt with in the same way. This, reduced to its essentials, is all that precedent means.

In civil-law nations all judicial decisions are, in theory, based upon legislative enactments, and the doctrine of judicial precedent does not apply. Practice, however, departs from theory. While there are comprehensive legislative codes in these countries, supposedly covering almost every aspect of human conduct and supplying ready-made answers for all problems that can arise, in fact many of the provisions are exceedingly vague and are sometimes almost meaningless until applied to concrete situations, when judicial interpretation gives them specific meaning. Furthermore, the legislative codes cannot anticipate all situations that may arise and come before the courts. The gaps in legislative schemes must be and are filled by judicial decisions, for no court in any nation is likely to refuse to decide a case on the ground that it has not been told in advance the answers to the questions presented to it. Decisions dealing with circumstances unforeseen by the codes and giving specific meaning to vague legislative provisions are published in most civil-law countries and are frequently referred to by lawyers and relied upon by judges. They are not considered "binding," but neither are they forgotten or disregarded. In actual practice, they have almost as much influence as statutory interpretations in nations that formally adhere to the doctrine of stare decisis.

It remains true that in common-law countries judicial lawmaking is more pervasive and more frankly acknowledged than in civil-law countries. In addition to rendering decisions that authoritatively interpret statutes, the courts of these nations have created a vast body of law without any statutory foundation whatever. Centuries ago, when there was no legislation to guide them, judges began to decide cases in accordance with their own conceptions of justice. Later judges followed them, deciding like cases in the same manner but distinguishing earlier cases when dissimilar factors were discovered in the cases before them. The later cases also became precedents to be followed in still later cases presenting substantially similar fact patterns. So the process has continued over centuries and is still continuing. The total accumulation of all these judicial decisions is what constitutes "the common law"--the by-product of judges deciding cases and setting forth their reasons. In the common-law nations, legislation is, as a result, more

limited in scope than in the civil-law countries. It does not purport to provide for all possibilities but leaves large areas of conduct to be governed solely by judge-made law.

To speak of precedent as "binding" even in common-law systems is misleading. As already noted, earlier decisions can be and are distinguished when judges conclude that they are based upon situations different from those before the court in later cases. Even more significant, earlier decisions can be overruled by the courts that rendered them (not by courts lower in the judicial hierarchy) when the judges conclude that they have proved to be so erroneous or unwise as to be unsuited for current or future application. The Supreme Court of the United States has overruled many of its own earlier decisions, to the consternation of those who yearn for a rigid separation of powers and who are unable to accept the inevitability of judicial lawmaking. Many of these overrulings are in the field of constitutional law, in which legislative correction of an erroneous judicial interpretation of the Constitution is impossible and in which the only alternative is the exceedingly slow, cumbersome, costly, and difficult process of constitutional amendment. Nevertheless, the power to overrule decisions is not restricted to constitutional interpretations. It extends to areas of purely statutory and purely judge-made law as well, areas in which legislative action would be equally capable of accomplishing needed changes. Even in England, which has no written constitution and which has traditionally followed a far more rigid doctrine of stare decisis than the United States, the House of Lords, in its role as the highest court, has announced its intention of departing from precedent "in appropriate cases."

The desirability of judicial lawmaking has long been the subject of lively debate in both civil- and common-law countries. That courts should not arrogate to themselves unrestricted legislative power is universally accepted. But when existing statutes and precedents are outmoded or barbarous as applied to specific cases before the courts, should not judges be able to change the law in order to

achieve what they conceive to be just results or, stated differently, to avoid what they consider unjust results?

The extent to which the judges should be bound by statutes and case precedents as against their own ethical ideas and concepts of social, political, and economic policy is an important question, as is the matter of which should prevail when justice and law appear to the judges to be out of alignment with each other. These are questions upon which reasonable persons disagree vigorously even when they are in basic agreement on the proposition that some degree of judicial lawmaking is inevitable. What is mainly at issue is the proper tempo and scope of judicial change. How quickly should judges act to remedy injustice and when should they consider an existing rule to be so established that its alteration calls for constitutional amendment or legislative enactment rather than judicial decision? As many dissenting opinions attest, judges themselves disagree on the answers to these questions, even when they are sitting on the same bench hearing the same case.

CONSTITUTIONAL DECISIONS

In some nations courts not only interpret legislation but determine its validity and in so doing sometimes render statutes inoperative. This happens only in nations that have written constitutions and have developed a doctrine of "judicial supremacy." The prime example is the United States, and the classic statement of the doctrine is the Supreme Court's decision in Marbury v. Madison (1803), in which Chief Justice Marshall said:

The powers of the legislature are defined and limited; and that those limits may not be mistaken, or forgotten, the Constitution is written. To what purpose are powers limited, and to what purpose is that limitation committed to writing, if these limits may, at any time, be passed by those intended to be restrained? The distinction between a government with limited and unlimited powers, is abolished, if those limits do not confine the persons on whom they are imposed, and if acts prohibited and acts allowed, are of equal obligation. It is a proposition too plain to be contested, that the Constitution controls any legislative act repugnant to it. It is

emphatically the province and duty of the judicial department to say what the law is. Those who apply the rule to particular cases, must of necessity expound and interpret that rule. If two laws conflict with each other, the courts must decide on the operation of each.

Armed with the authority asserted at this early date, the Supreme Court of the United States has held many statutes, federal as well as state, unconstitutional and has also invalidated executive actions that violated the Constitution. Even more surprising is the fact that lower courts also possess and exercise the same powers. Whenever a question arises in any U.S. court at any level as to the constitutionality of a statute or executive action, that court is obligated to determine its validity in the course of deciding the case before it. The case may have been brought for the sole and express purpose of testing the constitutionality of the statute or it may be an ordinary civil or criminal case, in which a constitutional question incidental to the main purpose of the proceeding is raised. Of course, when a lower court decides a constitutional question, its decision is subject to appellate review, sometimes at more than one level. When a state statute is challenged as violating the state constitution, the final authority is the supreme court of that state; when a federal or state statute or a state constitutional provision is challenged as violating the Constitution of the United States, the ultimate arbiter is the Supreme Court of the United States.

In a few American states, questions as to the constitutional validity of a statute may be referred in abstract form to the state's highest court by the chief executive or the legislature for an advisory opinion. This, however, is unusual and, in any event, supplementary to the normal procedure of raising and deciding constitutional questions. The normal pattern is for a constitutional question to be raised at the trial-court level in the context of a genuine controversy and to be decided finally on appellate review of the trial-court decision.

The U.S. pattern of constitutional adjudication is not followed in all nations that have written constitutions. In some, such as Germany, there is a special court at the highest level of government that handles only constitutional questions and to which

all such questions are referred as soon as they arise. A constitutional question may be referred to the special court in abstract form for a declaratory opinion by a procedure similar to that prevailing in the minority of U.S. states that allow advisory opinions.

In other nations, written constitutions may be in effect but not accompanied by any conception that their authoritative interpretation is a judicial function. Legislative bodies, rather than courts, act as the guardians and interpreters of the constitution, being guided by their provisions but not bound by them in any realistic sense.

Finally, there are some nations, such as England, that have no written constitutions. Here parliamentary supremacy clearly prevails. The courts have no power to invalidate statutes, although they can and do interpret them.

PROCEDURAL RULE MAKING

Distinct from the type of lawmaking just described is a more conscious and explicit type of judicial legislation and one that is less controversial. It is directed toward the rules of procedure by which the courts operate. This is a technical area in which expert knowledge of the type possessed by judges and lawyers is needed; in which constant attention to detail is required; and in which major problems of social, economic, or political policy are seldom encountered. Some legislative bodies, able or willing to devote only sporadic attention to the day-to-day problems of the management of litigation, have delegated the power to regulate procedure to the courts themselves. This is not ad hoc judicial lawmaking as a by-product of deciding cases but openly acknowledged promulgation of general rules for the future, in legislative form, by courts rather than legislatures.

An outstanding example of judicial rule making is found in the United States, where Congress has delegated to the Supreme Court broad power to formulate rules of civil, criminal, and appellate procedure for the federal courts. The Supreme Court also has and exercises the power to amend the rules from time to time as experience indicates that changes are desirable. Congress reserves the power to veto the rules so promulgated but has felt no need to exercise it.

Other legislative bodies, including those of some American states and most of the nations of continental Europe, have been unwilling to repose equal trust in the courts and have retained for themselves the power to regulate procedure. The results have been varied. Courts sometimes become so immersed in day-to-day decision making that they fail to pay adequate attention to the proper functioning of the judicial machinery and so perpetuate rules that are unduly rigid, unrealistic, and unsuited to the needs of litigants, which was the case in England and the American colonies during the 18th and first part of the 19th century. When such a condition occurs, reform through legislative action is indicated. Apart from the occasional necessity of major sweeping changes, however, experience in the common-law countries, at least, indicates that procedural rule making is better vested in the courts than in legislative bodies.

REVIEW OF ADMINISTRATIVE DECISIONS

Existing alongside the courts in any nation are administrative agencies of various kinds. Some do substantially the same kind of work as is done by courts and in substantially the same manner; some have quite different functions such as the issuing of licenses and the payment of welfare benefits.

The relationship between such agencies and regular courts differs markedly between common-law countries and civil-law countries. In common-law countries the actions of administrative agencies are subject to review in the ordinary courts. If the agency is one that decides controversies in substantially the same manner as a court, but in a different and more limited area, judicial control takes much the same form of appellate review as is provided for the decisions of lower courts. The objective of reviewing the record of proceedings is to determine whether the administrative agency acted within the scope of its jurisdiction, whether there was any evidence to support its conclusion, and whether the governing law was correctly interpreted and applied. Administrative decisions are seldom upset by the courts because of a belief on the part of most judges that administrative agencies have special expertise in the area of their specialty. However, they can be and occasionally are upset, thus underscoring the large degree of judicial control over

other agencies of government that characterizes common-law systems. If the administrative agency does not engage in formal adjudication, it produces no record of proceedings for judicial review. Nevertheless, its action can be challenged in court by way of trial rather than appeal. The same problems are presented for judicial determination: did the agency act within its jurisdiction, did it correctly follow the law, and was there any rational or factual basis for its action?

In many civil-law countries, the ordinary courts have no control over administrative agencies. Their decisions are reviewed by a special tribunal that is engaged exclusively in that work and that has nothing to do with cases of the type that come into the courts. Its function is solely appellate and solely within the specialized areas entrusted to the administrative agencies. The prototype of this type of tribunal is the Conseil d'État of France.

ENFORCEMENT OF JUDICIAL DECISIONS

The method of enforcing a judicial decision depends upon its nature. If it does nothing more than declare legal rights, as is true of a simple divorce decree (merely severing marital ties, not awarding alimony or the custody of children), or a declaratory judgment (for example, interpreting a contract or a statute), no enforcement is needed. If a judgment orders a party to do or refrain from doing a certain act, as happens when an injunction is issued, the court itself takes the first step in enforcing the judgment by holding in contempt anyone who refuses to obey its order and sentencing him to pay a fine or go to jail. Thereafter, enforcement is in the hands of the executive branch of government, acting through its correctional authorities.

In routine criminal cases and in civil cases that result in the award of money damages, courts have little to do with the enforcement of their judgments. That is the function of the executive branch of government, acting through sheriffs, marshals, jailers, and similar officials. The courts themselves have no machinery for enforcement.

Some judgments are extremely controversial, as was the case with the decision of the Supreme Court of the United States ordering racial desegregation of the schools. When voluntary compliance with such a judgment is refused, forcible methods of enforcement are necessary, sometimes extending to the deployment of armed forces under the control of the executive branch of the government. The withdrawal of executive support seldom occurs, even when decisions are directed against the executive branch itself; but when such executive support is withheld, the courts are rendered impotent. Judges, being aware of their limited power, seldom render decisions that they know to be so lacking in support that they will not be enforced.

Court structure and organization

TYPES OF COURTS

There are many different types of courts and many ways to classify and describe them. Basic distinctions must be made between civil and criminal courts, between courts of general jurisdiction and those of limited jurisdiction, and between trial and appellate courts.

Criminal courts.

Criminal courts deal with persons accused of crime, deciding whether they are guilty and, if so, determining the consequences they shall suffer. Prosecution is on behalf of the public, represented by some official such as a district attorney, procurator, or a police officer.

Courts are also public agencies, but in this instance they stand neutral between the prosecution and the defense, their objective being to decide between the two in accordance with law.

In civil-law countries a more active role is assigned to the judge and a more passive role to counsel than in common-law countries. In the common-law courts, in which the "adversary" procedure prevails, the lawyers for both sides bear responsibility for producing evidence and they do most of the questioning of witnesses. In civil-law countries, "inquisitorial" procedure prevails, with judges doing most of the questioning of witnesses and having an independent

responsibility to discover the facts. This difference pertains more to procedure rather than function.

If a person has been found guilty, he is sentenced, again according to law and within limits fixed by legislation. The objective is not so much to wreak vengeance upon the offender as to rehabilitate him and deter others from following his example. Hence the most common sentences are fines, short terms of imprisonment, and probation (which allows the offender to remain at large but under supervision). In extremely serious cases, the goal may be to prevent the offender from committing further crimes, which may call for a long term of imprisonment or even capital punishment. The death penalty, however, is gradually disappearing from the criminal codes of civilized nations.

Criminal proceedings in any nation inevitably have some educational impact on defendants and on members of the general public. In Communist nations education is a conscious and primary goal. A basic provision of old Soviet law declared:

By all its activities the court shall educate the citizens of the U.S.S.R. in the spirit of devotion to the Motherland and the cause of communism in the spirit of strict and undeviating observance of Soviet laws, of care for socialist property, of labor discipline, of honesty toward public and social duty, of respect for the rights, honor and dignity of citizens, for the rules of socialist common-life.

Civil courts.

Civil courts deal with "private" controversies, as where two individuals (or corporations) are in dispute over the terms of a contract or over who shall bear responsibility for an auto accident. Ordinarily the public is not a party as in criminal proceedings, for it has no interest beyond providing just rules for decision and a forum where the dispute can be impartially and peacefully resolved.

It is possible, however, for the government to be involved in civil litigation if it stands in the same relation to a private party as another individual might stand. Thus, if a postal truck should run down a pedestrian, the government might be sued civilly by the injured person; or if the government contracted to purchase supplies that turned out to be defective, it might sue the dealer for damages in a civil court.

The objective of a civil action is not punishment or correction of the defendant or the setting of an example to others but rather to restore the parties so far as possible to the positions they would have occupied had no legal wrong been committed. The most common civil remedy is a judgment for money damages, but there are others, such as an injunction ordering the defendant to do, or refrain from doing, a certain act or a judgment restoring property to its rightful owner.

Civil claims do not ordinarily arise out of criminal acts. A person who breaks his contract with another or who causes him a physical injury through negligence may have committed no crime but only a civil wrong for which he may not be prosecuted criminally by the public.

There are, however, areas of overlap, for a single incident may give rise to both civil liability and criminal prosecution. In some nations, such as France, both types of responsibility can be determined in a single proceeding under a concept known as adhesion by which the injured party is allowed to assert his civil claim in the criminal prosecution, agreeing to abide by its outcome. This removes the necessity of two separate trials. In common-law countries there is no such procedure, even though civil and criminal jurisdiction may be merged in a single court. Two separate actions must be brought, independent of each other.

Courts of general jurisdiction.

Although there are some courts that handle only criminal cases and others that handle only civil cases, a more common pattern is for a single court to be vested with both civil and criminal jurisdiction. Such is the High Court of England and such are many of the trial courts found in U.S. states. Often these tribunals are called courts of general jurisdiction, signifying that they can deal with almost any type of controversy, although in fact they may not have jurisdiction over certain types of cases assigned to specialized tribunals. Often such courts are also described as superior courts, because they are empowered to handle serious criminal cases and important civil cases involving large amounts of money.

Even if a court possesses general or very broad jurisdiction, it may nevertheless be organized into specialized branches, one handling criminal cases, another handling

civil cases, another handling juvenile cases, and so forth. The advantage of such an arrangement is that judges can be transferred from one type of work to another, and cases do not fail to be heard for having been instituted in the wrong branch since they can be transferred administratively with relatively little effort.

Courts of limited jurisdiction.

Specialized tribunals of many kinds exist, varying from nation to nation. Some deal only with the administration of the estates of deceased persons (probate courts), some only with disputes between merchants (commercial courts), some only with disputes between employers and employees (labour courts). All are courts of limited jurisdiction. Deserving of special mention because of their importance are juvenile courts, empowered to deal with misconduct by children and sometimes also with the neglect or maltreatment of children. Their procedure is much more informal than that of adult criminal courts, and the facilities available to them for the pretrial detention of children and for their incarceration, if necessary after trial, are different. The emphasis is on salvaging children, not punishing them.

Traffic courts also deserve mention because they are so common. They process motor vehicle offenses such as speeding and improper parking. Their procedure is summary and their volume of cases heavy. Contested trials are relatively infrequent.

Finally, in most jurisdictions there are what are called, unfortunately and for want of a better term, "inferior" courts. These are often manned by part-time judges who are not trained in the law. They handle minor civil cases involving small sums of money, such as bill collections, and minor criminal cases carrying light penalties, such as simple assaults. In addition to finally disposing of minor criminal cases, such courts may handle the early phases of more serious criminal cases--fixing bail, advising defendants of their rights, appointing counsel, and conducting preliminary hearings to determine whether the evidence is sufficient to justify holding defendants for trial in higher "superior" courts.

Appellate courts.

The tribunals described thus far are trial courts or "courts of first instance." They see the parties, hear the witnesses, receive the evidence, find the facts, apply the law, and determine the outcome.

Above them, to review their work and correct their errors, are appellate courts. These are usually collegiate bodies, consisting of several judges instead of the single judge who usually presides over a trial court. The jurisdiction of the appellate courts is usually general; specialized appellate tribunals handling, for example, only criminal appeals or only civil appeals are rare, although not unknown. The Conseil d'État of France and the Federal Constitutional Court of Germany have already been mentioned as examples of specialization.

Appellate review is not automatic. It must be sought by some party aggrieved by the judgment in the court below. For that reason, and because an appeal may be both expensive and useless, there are far fewer appeals than trials and, if successive appeals are available, as is often the case, far fewer second appeals than original appeals. Judicial systems are organized on a hierarchical basis: at the bottom are numerous trial courts scattered throughout the nation; above them are a smaller number of first-level appellate courts, usually organized on a regional basis; and at the apex is a single court of last resort.

There are three basic types of appellate review. The first consists of a retrial of the case, with the appellate court hearing the evidence for the second time, making fresh findings of fact, and in general proceeding in much the same manner as the court that originally rendered the judgment under attack. This "trial de novo" is used in common-law countries for the first stage of review but only when the trial in the first instance was conducted by an "inferior" court--one typically manned by a part-time judge or two or more such judges, empowered to try only minor cases and keeping no adequate record of its proceedings.

The second type of review is based in part on a "dossier," which is a record compiled in the court below of the evidence received and the findings made there. The reviewing court has the power to rehear the same witnesses again or to supplement their testimony by taking additional evidence, but it need not and

frequently does not do so, being content to rely on the record already made in reaching its own findings of fact and conclusions of law. This type of proceeding prevails generally in civil-law countries for the first stage of appellate review, even when the original trial was conducted in a superior court, staffed by professional judges, and empowered to try important or serious cases.

The third type of review is based solely on a written record of proceedings in the court or courts below. The reviewing court does not itself receive evidence directly but concentrates its effort on discovering from the record whether any errors were committed of such a serious nature as to require reversal or modification of the judgment under attack or a new trial in the court below. The emphasis is on questions of law (both procedural and substantive) rather than on questions of fact. This type of review prevails both in civil-law nations and common-law nations at the highest appellate level. It is also used in common-law nations at lower levels when the judgment of a superior court is under attack. The purpose of this type of review is not merely to assure that correct results are reached in individual cases but also to clarify and expound the law in the manner described earlier. Lower courts have little to do with the development of the law, for they ordinarily do not write or publish opinions. The highest appellate courts do, and it is their opinions that become the guidelines for future cases.

Courts in federal systems.

Many nations, such as England, France, and Japan, have unitary judicial systems with all courts (that is, regular courts as distinguished from administrative bodies) fitting into a single national hierarchy of tribunals along the lines just described. Other nations, organized on a federal basis, tend to have more complicated court structures, reflecting the fragmentation of governmental powers between the central authority and the local authorities. In the United States, for example, there are 51 separate judicial systems, one for each state and another for the federal government. To a limited extent, the jurisdiction of the federal courts is exclusive of that exercised by the state courts, but there are large areas of overlap and duplication. At the top level is the Supreme Court of the United States, hearing

appeals not only from the lower federal courts but also from state courts insofar as they present federal questions arising under the Constitution of the United States or under federal statutes or treaties. If a case in a state court involves only a question of state law--for example, the interpretation of a state statute--the ultimate authority is the state supreme court, and no appeal is possible to the Supreme Court of the United States.

Court structure in a federal form of government need not be as complicated as that in the United States. It is possible to have only one set of courts for the nation, operated by the central government and handling all cases that arise under state law as well as federal law.

Another possibility is for each state or province to have its own system of courts, handling all questions of federal as well as state law, and for the central government to maintain only a single supreme court to decide questions as to the relationship of the central authority and the local authorities or as to the relationship among the local authorities themselves. This is the pattern in Canada and Australia.

Another complication resulting from a federal form of government is that questions involving conflict of laws arise with great frequency. Such questions concern the choice to be made between the law of one jurisdiction and another as the rule for decision in a particular case.

Even in a unitary system, such problems cannot be avoided, for an English court may be called upon to try a case arising from a transaction that took place in France and to decide whether English or French law should govern. Such problems arise much more often, however, in federal systems, where laws differ from state to state and people move about very freely. Their activities in one state sometimes become the subject of a lawsuit in another, requiring the court to decide which law should apply.

JUDGES

A court is a complex institution whose functioning depends upon many people: not only the judge but also the parties, their lawyers, witnesses, clerks, bailiffs,

probation officers, administrators, and many others, including, in certain types of cases, jurors. Nevertheless, the central figure in any court is the judge.

Judges vary enormously, not only from nation to nation but often within a single nation. For example, a rural justice of the peace in the United States--untrained in the law, serving part-time, sitting alone in work clothes in a makeshift courtroom, collecting small fees or receiving a pittance for salary, trying a succession of routine traffic cases and little else--obviously bears little resemblance to a justice of the Supreme Court of the United States--a full-time, well-paid, black-robed professional, assisted by law clerks and secretaries, sitting in a marble palace with eight colleagues and deciding at the highest appellate level only questions of profound national importance. Yet both persons are judges.

Lay judges.

In some civil-law countries, judges at all levels are professionally trained in the law, but in many other nations they are not. In England, part-time lay judges outnumber full-time professional judges by about 60 to 1. Called magistrates or justices of the peace, they dispose of about 97 percent of all criminal cases in that nation and do so with general public satisfaction and the approbation of most lawyers. Professional judges deal only with the most serious crimes, which are relatively few in number; most of their time is devoted to civil cases. England places unusually heavy reliance on lay judges, but they are far from unknown in the courts of many other nations, particularly at the lowest trial level. This was also true in the former U.S.S.R. and remains so in the United States. There is considerable diversity in the way laymen are chosen and used in judicial work. In the United States, for example, lay judges are popularly elected for limited terms, whereas in England they are appointed by the lord chancellor to serve until retirement or removal. In England the lay judges serve intermittently in panels on a rotating basis for short periods, whereas in the United States they sit alone and continuously. In the erstwhile U.S.S.R. lay judges (called assessors) always sat with professional judges; in England, they sometimes do; and in the United States, they never do. In some underdeveloped nations, few judges at any level are legally

trained. They are more likely to be priests, for the law they administer is mainly derived from religious teaching, and religion and secular government are often not sharply differentiated. The vast majority of nations that use lay judges at the lowest trial level, however, insist upon professionally trained judges at higher levels: in trial courts of general jurisdiction and in appellate courts.

Professional judges in the civil-law tradition.

Professional judges in civil-law countries are markedly different in background and outlook from professional judges in common-law countries. Both are law-trained and both perform substantially the same functions, but there the similarities cease. In a typical civil-law country, a person graduating from law school makes a choice between a judicial career and a career as a private lawyer. If he chooses the former and is able to pass an examination, he is appointed to the judiciary by the minister of justice (a political officer) and enters service in his early 20s. His first assignment is to a low-level court; thereafter, he works his way up the judicial ladder as far as he can until his retirement on a pension. His promotions and assignments depend upon the way his performance is regarded by a council of senior judges, or sometimes upon the judgment of the minister of justice, who may or may not exercise his powers disinterestedly and on the basis of merit. The civil-law judge, in short, is a civil servant.

Professional judges in the common-law tradition.

In common-law nations, the path to judicial office is quite different. Upon completion of his formal education, a person typically spends 15, 20, or 25 years in the private practice of law or, less commonly, in law teaching or governmental legal service and then, at about age 50, becomes a judge. He takes no competitive examination but is appointed or elected to office. In England the appointive system prevails for all levels of judges, including even lay magistrates. Appointments are primarily under the control of the lord chancellor, who, although a cabinet officer, is also the highest judge of the realm. They are kept surprisingly free from party politics. In the United States, the appointive method is used in federal courts and some state courts, but it tends to be highly political. Appointments are made by the

chief executive of the nation or state and are frequently subject to legislative approval. In many states, judges are popularly elected, sometimes on nonpartisan ballots, sometimes on partisan ballots with all the trappings of traditional political contests. In an attempt to de-emphasize political considerations and yet maintain some measure of popular control over the selection of judges, a third method of judicial selection has been devised and is slowly growing in popularity. Called the Missouri Plan, it involves the creation of a nominating commission that screens judicial candidates and submits to the appointing authority a limited number of names of persons considered qualified. The appointing authority must make his choice from the list submitted. The person chosen as judge then assumes office for a limited time, and, after the conclusion of this probationary period, he stands for "election" for a much longer term. He does not run against any other candidate but only "against his own record."

In common-law countries, a person does not necessarily enter the judiciary at a low level; he may be appointed or elected to his nation's highest court or to one of its intermediate courts. He does not look forward to any regular pattern of promotion, nor is he necessarily assured of long tenure with ultimate retirement on a pension. In some courts, life tenure is provided, usually subject to mandatory retirement at a fixed age. In others, tenure is limited to a stated term of years. At the conclusion of his term, if not mandatorily retired earlier, the judge must be reelected or reappointed if he is to continue.

While in office, the common-law judge enjoys greater power and prestige and more independence than his civil-law counterpart. He occupies a position to which most members of his profession aspire. He is not subject to outside supervision and inspection by any council of judges or by a minister of justice; nor is he liable to be transferred by action of such an official from court to court or place to place. The only administrative control over him is that exercised by judicial colleagues, whose powers of management are generally slight, being limited to such matters as requiring periodical reports of pending cases and arranging for temporary (and usually consensual) transfers of judges between courts when factors such as illness

or congested calendars require them. Only if a judge misbehaves very badly is he in danger of disciplinary sanctions and then usually only by way of criminal prosecution for his misdeeds or legislative impeachment and trial, resulting in removal from office--a very cumbersome, slow, ill-defined, inflexible, ineffective, and seldom used procedure. In parts of the United States, newer and more expeditious methods of judicial discipline are developing in which senior judges are vested with power to impose sanctions ranging from reprimand to removal from office of erring colleagues. They are also vested with power to retire judges who have become physically or mentally unfit to discharge their duties.

Except at the very highest appellate level, common-law judges are no less subject than their civil-law counterparts to appellate reversals of their judgments. But appellate review cannot fairly be regarded as discipline. It is designed to protect the rights of litigants; to clarify, expound, and develop the law; and to help and guide rather than reprimand lower court judges.

JURIES

The jury is a historic legal institution in which a group of laymen participate in a major way in deciding cases brought to trial. Its exact characteristics and powers depend on the laws and practices of the countries, provinces, or states in which it is found, and there is considerable variation. Basically, however, it recruits laymen at random from the widest population for the trial of a particular case and allows them to deliberate in secrecy, to reach a decision by other than majority vote, and to make it public without giving reasons.

History and use.

The jury's origin is lost in the past. It may have been indigenous to England or have been brought there by the Norman invaders in 1066. Originally, the jurors were neighbourhood witnesses who passed judgment based on what they themselves knew. But the breakdown of medieval society and the growth of the towns changed this; the jury was called upon to determine the facts of the case, based upon the evidence presented in court. The availability of the jury in the king's courts may have been a key factor in centralizing the nation's courts under

the king and in creating the common law. By the 15th century, nonrational modes of trial such as ordeal, in which the defendant was subjected to various tortures that, if successfully endured, proved his innocence, were replaced by the jury trial, which became the established form of trial for both criminal and civil cases at common law.

Two forces moved the jury abroad. One was the expansion of the British Empire, which brought the jury to Asia, Africa, and the American continent. The other was the French Revolution and its aftermath, which brought it, as a symbol of popular government, to the European continent: first to France itself, then, through Napoleon, to the Rhineland, later to Belgium, most of the remaining German states, Austria-Hungary, Russia, Italy, Switzerland, Holland, and Luxembourg, although the last two abolished it immediately after Napoleon's defeat. In each of these countries, use of the jury was from the outset limited to trials of major crimes and of political crimes against the state.

Beginning in the mid-19th century, the jury was weakened in a variety of ways: in 1850, Prussia, for example, removed treason from its jurisdiction; in 1851, the duchy of Nassau removed all political crimes; in 1923, Czechoslovakia removed treason and, one year later, libel; in 1919, Hungary suspended jury trial entirely and never restored it. Germany abandoned the jury in 1924. Both the Soviet bloc and the fascist states abolished it outright; France never restored the jury abolished during the German occupation in the 1940s, and Japan did away with its short-lived jury courts in 1943. After World War II, Austria reintroduced the jury in a weakened form.

Thus, there are three important points about the history and development of the jury as a legal institution: first, the effort to introduce it outside the Anglo-American legal orbit has failed; further, in England itself its use was limited by statute to a small category of cases; and, thus, the United States has emerged today as the home of the jury system for both criminal and civil cases. Some 120,000 jury trials are conducted there annually, more than 90 percent of all jury trials in the world.

Use of the jury in the United States depends on two factors: the degree to which it is available as a matter of right and the degree to which the parties themselves choose to use it. The laws as to its availability have varied from state to state, but in 1968 in Duncan v. Louisiana the United States Supreme Court declared that a jury trial is a constitutional right in all criminal cases in which the penalty may exceed six months' imprisonment. In civil cases its constitutional status is less clear, but, in general, jury trial is available. The practice of allowing the parties to waive a jury trial also varies widely from region to region, and, as a result, the number of jury trials per year also varies widely. The annual number of criminal jury trials per 100,000 population ranges between 3 for Connecticut to 144 for Georgia.

Jury procedures.

Selection.

Historically there were some minimum requirements of property and competence for jury service. More recently the idea of genuine random selection from the population, to achieve a cross section of the community, has been gaining ground. Since 1969 it has been the principle of selection in the federal courts. Most jurisdictions exempt some groups from jury service: police officers, lawyers, doctors, and so on. All jurisdictions excuse jurors if the service imposes undue hardship.

The commitment of important decisions to a random group of laypersons has been moderated, particularly in the United States, by an elaborate screening, voir dire, conducted by trial counsel at the inception of a trial. The law permits counsel to challenge prospective jurors either for cause (if there is specific likelihood of bias) or, for a limited number, "peremptorily"--that is, without having to give a reason. American trial tradition attaches a great deal of significance to the strategies of juror selection, and in celebrated cases the lawyers' voir dire examination has extended for several weeks.

Size and unanimity.

Traditionally the jury had 12 members and was required to reach its decisions with unanimity, a striking arrangement in Anglo-American countries that make all other decisions by majority vote. Over the years some modifications have been made. Some jurisdictions prescribe or allow in minor cases a jury of six. Oregon allows 10:2 verdicts--that is, a majority of 10--in all criminal cases, except capital ones; and in 1968 England followed the Oregon example. A few Southern states in the United States allow majority verdicts in misdemeanour trials. In civil cases many states now allow 10:2 verdicts. When the required number (12 or 10) of jurors cannot agree on a verdict (termed a hung jury in the United States), the judge declares a mistrial, which means the case, unless it is withdrawn, must be tried anew. It is somewhat remarkable that "hung" juries occur with relative infrequency even when unanimity is required. In Europe juries operate under a different principle. Unless at least two-thirds of all the jurors vote guilty, the defendant must be acquitted. The U.S. Army court-martial jury also operates under this principle.

Sentencing.

Although in civil cases the jury decides both issues of liability and amount of damages, in criminal cases it has been restricted generally to the issues of guilt, while punishment has been left to the judge. In some Southern U.S. states, however, the jury also decides the sentence within a certain range that the law provides. And, in all jurisdictions that have retained the death penalty, if the jury finds the defendant guilty of the capital crime, it decides, or at least expresses an opinion, as to whether the death penalty is to be imposed. In most jurisdictions decisions on guilt and sentence are rendered simultaneously, but California has introduced the so-called second trial in capital cases, which occurs after a guilty verdict. At such a "second trial" pleas and evidence are presented for and against the imposition of the death penalty in the specific case, and only then is the jury asked to determine the sentence.

Control.

Trial by jury is, of course, trial by jury under the supervision of a judge. The formula for sharing power between judge and jury is complex. First, the judge

decides what the jury may or may not hear under the rules of evidence. Second, if the judge finds that the evidence presented leaves no factual issue to be resolved, he may withdraw the issue from the jury and direct the jury to acquit a defendant or, in a civil trial, find for either plaintiff or defendant; he cannot, however, direct a guilty verdict in a criminal trial. Third, in some jurisdictions the judge may, and often will, summarize the evidence or even discuss its weight. Fourth, the judge instructs the jury as to the law it should apply in reaching the verdict. Finally, if the judge finds the jury's verdict to be manifestly against the weight of the evidence, he may with one exception set it aside and order a new trial. The only exception is in a criminal case in which the jury renders an acquittal; under Anglo-American law (though not under European continental law) the jury's acquittal is always final.

The jury normally renders a general verdict--that is, a yes or no answer to liability or guilt--and does not give reasons for its decision. At times, however, courts employ "special verdicts" or "special interrogatories" in which the jurors are asked to decide a series of specific factual issues that bear on the overall verdict.

The controversy over the jury.

The jury has been enmeshed in a perennial debate as to its merits, a debate that has recruited some of the great names in law and political philosophy--from Montesquieu, William Blackstone, and Thomas Jefferson to present-day theorists and practitioners--and has centred on three issues. First, there is the debate about collateral aspects: there are favourable contentions that the jury provides an important civic experience, that it makes tolerable the stringency of certain decisions, that it acts as a sort of lightning rod for animosity that otherwise might centre on the more permanent judge, and that the jury is a guarantor of integrity since it is said to be more difficult to bribe 12 people than one. Against this it has been urged that jury duty disenchants the citizen, that it imposes an unfair burden, that the jury is expensive, and that it makes it difficult to do away with the often interminable delays that exist in civil litigation.

Second, there is the issue of the jury's competence. It is argued that the judge, by training, discipline, experience, and superior intelligence, is better able to understand law and facts than laypersons drawn from a broad range of levels of intelligence, without experience and without durable official responsibility. But it is also argued that 12 heads are better than one, that the jury as a group has wisdom and strength beyond that of its individual members, that it makes up in common sense and experience what it lacks in training, and that its very inexperience is an asset because it secures a fresh perception of each trial, avoiding the stereotypes that may infect the judicial eye.

Finally, there is the question of the jury's interpretation of the law. The critics complain that the jury will not follow the law, either because it does not understand it or because it does not like it, and hence will administer justice unevenly and that the jury produces a government by individuals and not by rule of law, against which Anglo-American political tradition is so steadfastly set. The jury's champions offer this very flexibility as its most endearing characteristic. They see the jury as a remarkable device for ensuring that the rigidity of the general rule can be shaped to justice in a particular case, with government by the spirit of the law and not by its letter.

Performance.

In a recent survey of some 7,000 jury trials, the presiding judges were requested to reveal how they would have decided without a jury; the results of the survey provided some major insights into the actual performance of the contemporary American jury. In both civil and criminal trials, judge and jury agreed in 78 percent of all verdicts. In civil cases the disagreement in the remaining cases was symmetrically split; in 19 percent of the criminal cases, however, the judge would have convicted, whereas the jury acquitted. The letter of the law confines the jury to "finding the facts," but the deviations from the judge are mostly due to the jury's subtle, and not always conscious, injecting its sense of justice into a case that might go either way. This sense of justice may be concerned with the person of the accused, with the threat of too harsh a punishment, or with the content of the

criminal law rules. Thus, close study of the jury has revealed it as a highly sensitive institution, subtle and discerning, moved by factors far beyond gross sympathy for the defendant. On the whole, the system tolerates and even appreciates these deviations of the jury from the judge, even if in rare cases they reflect what the national community experiences as intolerable local prejudice.

OTHER JUDICIAL OFFICIALS

In most countries there are other officials who serve the court. Court clerks, who are responsible for case records and documents, and bailiffs, who are in charge of keeping order, are found in most judicial systems. Also prevalent are officers who prosecute cases in the government's name: states attorneys and district attorneys in the United States, procurators-general in the U.S.S.R., and procureurs généraux in France.

Probation officers are found in many countries including the U.S. and Japan. Notaries in France, Italy, and the U.S.S.R. have greater powers than their counterparts in the U.S. In fact, they perform many services carried out by lawyers in the common-law system, such as drafting and verifying wills and contracts and preparing petitions for presentation in court.

Certain countries have officials that are particularly indigenous to their country or legal system. France, for example, has a juge d'instruction, who is responsible for the preliminary investigative proceedings prior to a criminal trial.

THE STRUCTURE AND STATUS OF THE JUDICIARY UNDER COMMUNISM

Although the essential legal institutions of the Soviet Union and other Communist countries are based on the civil-law system, certain features are unique. These characteristics are partly the result of the Soviet Union's attitudes toward law that antedate the Soviet system, but mostly they result from the attempt to reconcile Marxist theory with the institutional needs of a modern society.

According to Marx and his followers, the legal system, like all other governmental structures and instruments of class oppression, would "wither away" in a Communist society; thus the courts that existed after the Revolution were

considered temporary institutions, required only during the transition to Communism. The ordinary and traditional business of the courts was carried on by the so-called people's courts, while "revolutionary tribunals" dealt with individuals the government considered to be political opponents. A nonjudicial body in the hands of the secret police (at first called Cheka, later OGPU and NKVD), operating in the style of an administrative agency, also heard cases and handed out sentences--usually of the severest kind.

In 1921 some capitalist measures were temporarily introduced to revive the economy, and this necessitated some stabilization of the legal system and its institutions. A three-level system of courts with civil and criminal jurisdiction was established in 1922 for the Russian Republic, which in the same year formed a federation with the other soviet republics under its jurisdiction, making up the Union of Soviet Socialist Republics. A new constitution created a federal court--the U.S.S.R. Supreme Court--and a federal judiciary act of 1924 established uniform principles for the judiciary throughout the republics, patterned largely after the system adopted by the Russian Republic. The basic structure of the courts laid down at that time has remained essentially the same to the present, with some minor changes and reforms.

The "People's Courts" on the local level are courts of original jurisdiction for minor criminal cases and a large number of civil cases. The next level, the provincial courts, receive appeals from the people's courts and have original jurisdiction over political and serious civil and criminal cases. The highest level in each republic is its supreme court, which hears appeals from the provincial courts, disciplines lower courts, and has some original jurisdiction over extremely serious cases.

On all three levels, appellate cases are tried by a court consisting of three full-time judges, whereas one judge and two lay judges, or assessors, preside over cases on first hearing. Judges of the people's courts are popularly elected every five years, and judges on the provincial and supreme court levels are "elected" by soviets (bodies combining legislative and executive functions) of the corresponding levels

of government. All judges may be recalled before the expiration of their terms by those who elected them.

The federal court system is twofold. There are courts called military tribunals that deal with charges against men in the armed forces and with charges of espionage brought against civilians. The other federal body is the U.S.S.R. Supreme Court--the highest judicial body--which has original jurisdiction in a few special cases relating to the survival of the regime, appellate power over the decisions of the supreme courts of the republics or decisions of the military tribunals, and the right to issue directives to all inferior courts in matters of administration of justice on the basis of a series of its decisions. Although Soviet legal theory is patterned after that of civil-law countries in that it does not recognize judicial lawmaking, these Supreme Court directives function as a source of law and are binding on all courts.

The status of the judiciary in the Soviet Union has undergone some changes that parallel the institutional changes since the early days of the Revolution. The system organized in 1922 had the stated purpose of safeguarding the conquests of the Revolution and establishing the dictatorship of the proletariat. Judges were called upon to use their "revolutionary conscience" in deciding cases, and the doctrine of impartiality and independence of the judiciary was repudiated. With the passage of time, however, the Soviet rulers found the need for legal institutions of a stable nature increasing rather than decreasing, and the goal of the legal system was changed from protection of a particular class to protection of the socialist order and the rights of all citizens. Although the role of the judiciary is still conceived of as a political task, there is some acceptance of the idea that judges should be independent and impartial. Marxist philosophy notwithstanding, the Soviet Union and other Socialist countries are confronted with a growing need for legal institutions to fulfill many of the same functions as those in the West. One attempt to fill this need in recent years has been the appearance of "social organizations," such as the "comrades' courts," which are described as voluntary organizations using persuasion and social influence to deal with matters that would otherwise

come before a court. But these organizations are party controlled, have only limited power to impose sanctions, and do not appear at present to offer an effective alternative to the type of legal institutions that have been developing within Soviet society.

The other Communist countries, both in eastern Europe and Asia, adopted legal institutions patterned largely after the Soviet model. Since Stalin's death, however, there have been some modifications in the eastern European countries, coinciding with the reforms in the civil and criminal codes adopted by the Soviet Union in the late 1950s and early 1960s. Chinese leaders, however, have resisted efforts to codify their laws, preferring flexibility in their courts, and they have abandoned the policy of copying Soviet legal patterns.

Arbitration

Arbitration is a nonjudicial, legal technique for resolving disputes by referring them to a third party for a binding decision, or "award," as an arbitrator's findings are usually described. The arbitrator may be a single person or an arbitration board, usually of three members. Arbitration is most commonly resorted to for the resolution of commercial disputes and must be distinguished from mediation and conciliation, which are common in the settlement of labour disputes between management and labour unions. In such cases, the parties resort to a third person to offer a recommendation for a settlement or to help them to reach a compromise. Such intervention by a third party, which also occurs in international disputes between states in the form of diplomatic intervention and good offices, has no binding force upon the disputants, as has the arbitrator's decision, the award.

COMMERCIAL ARBITRATION

Commercial arbitration is a means of settling disputes by referring them to a third person, an arbitrator, selected by the parties for a decision based on the evidence and arguments presented to the arbitration tribunal. The parties agree in advance that the decision will be accepted as final and binding upon them.

Historically, commercial arbitration was used in resolving controversies between medieval merchants, in fairs and marketplaces in England and on the European

continent, and in the Mediterranean and Baltic sea trade. The increased use of commercial arbitration became possible after courts were empowered to enforce the parties' agreement to arbitrate. The first such statute was the English Arbitration Act of 1889, now consolidated into an act of 1950 and adopted by arbitration statutes in most countries of the Commonwealth. It was followed in the United States by an arbitration statute of the state of New York in 1920 and a Federal Arbitration Act of 1925. Codified in 1940, the latter deals with the enforcement in federal courts of arbitration agreements and awards in maritime transactions and those involving interstate and foreign commerce. Most states of the United States adopted, sometimes with minor changes, the Uniform Arbitration Act of 1955, as amended in 1956, which had been promoted by the Commissioners on Uniform State Laws and recommended by the American Bar Association. This act provides for the judicial enforcement of an agreement to arbitrate existing and future disputes, thereby making the arbitration agreement no longer revocable, as it had been under common law. It also provides for the substitution of arbitrators in the event of a party failing to select an arbitrator and for a suspension of any court action instituted in contravention of a voluntary arbitration agreement. The courts thereby play an important role in implementing arbitration agreements, making judicial assistance available against a recalcitrant party. This concept of modern arbitration law, which recognizes the irrevocability of arbitration agreements and the enforceability of awards prevails also in the arbitration statutes of nearly all countries of Europe and Asia. Latin American procedural laws generally provide only for court enforcement of agreements to arbitrate existing disputes and do not provide for the enforcement of subsequent disputes that may arise under the arbitration agreement.

Function and scope.

Arbitration has been used customarily for the settlement of disputes between members of trade associations and between different exchanges in the securities and commodities trade. Form contracts contain a standard arbitration clause referring to the arbitration rules of the respective organization. Numerous

arrangements between parties in industry and commerce also provide for arbitration of controversies arising out of contracts for the sale of manufactured goods, for terms of service of employment, for construction and engineering projects, for financial operations, for agency and distribution arrangements, and for many other undertakings.

The usefulness and significance of arbitration is demonstrated by its increasing use by the business community and the legal profession in many countries of the world. The primary advantage is the speed with which controversies can be resolved by arbitration, compared with the long delays of ordinary court procedure. The expert knowledge of arbitrators of the customs and usages of a specific trade makes testimony by others and much documentation unnecessary, thereby eliminating expenses connected with court procedures. The privacy of the arbitration procedure is also much valued by parties to the controversy; situations unfavourable to the party's credit or deficiencies in manufactured goods revealed in arbitration proceedings do not become known to outsiders. There are, however, also disadvantages in the arbitration process. The fact that in Anglo-American practice no reasons are given by the arbitrator to accompany his award prevents the development of a guideline for the further conduct of business relations. This uncertainty resulting from lack of reasoned precedents, moreover, makes the arbitral decision less predictable. Further obstacles to the wider use of commercial arbitration are the divergencies in municipal laws and court decisions that result in different interpretations of similar arbitration questions and the fact that awards are not published: publication of awards, even without identification of the parties, might assist in the establishment of precedents useful in discouraging future disputes on similar issues in a specific branch of industry or commerce.

Procedure.

The method of selecting arbitrators is an important aspect of the arbitration process, for the arbitrator's ability and fairness is the decisive element in any arbitration. The general practice is for both parties to select an arbitrator at the time a conflict arises or at the time the arbitration agreement is concluded. The two

arbitrators then select a chairman, forming a tribunal. Selection of arbitrators is also often made by agencies administering commercial arbitration under preestablished rules of procedure. These organizations--various trade associations, produce exchanges, and chambers of commerce in many countries--maintain panels of expert arbitrators. The parties may either make their own selection or entrust the appointment of the arbitrators to the organization.

Challenges to the arbitration process are not uncommon. A party may claim, for example, that no valid arbitration agreement came into existence because the person signing the agreement had no authority to do so or that a condition precedent to arbitration had not been fulfilled. More often, arbitration is contested on the ground that the specific controversy is not covered by the agreement. In such cases, the issue of whether the arbitrator has authority to deal with the conflict is usually determined by a court. Further challenges to the arbitration process may be directed against an arbitrator, on grounds, for example, of alleged lack of impartiality. Any such challenge can generally be maintained only after the arbitration has been concluded, as courts are reluctant to interfere with the arbitration process before an award has been rendered.

The arbitration process is governed by the rules to which the parties referred in their agreement; otherwise, the procedure will be determined by the arbitrators. The arbitration proceedings must be conducted so as to afford the parties a fair hearing on the basis of equality.

The arbitrator generally has the authority to request the parties and third persons to produce documents and books and to enforce such a request by issuing subpoenas. If a party fails to appear at a properly convened hearing, without showing a legitimate cause, the arbitrator in most instances will proceed in the absence of that party and render an award after investigation of the matter in dispute.

Under the law and arbitration practice of most countries, the award is valid and binding upon the parties when rendered by a majority of the arbitrators, unless the parties expressly request a unanimous decision of the arbitrators, which they seldom do. The statutory law of various countries and the rules of agencies

administering commercial arbitration contain provisions on the form, certification, notification, and delivery of the award, with which requirements the arbitrator has to comply.

A much-disputed question in commercial arbitration concerns the law to be applied by the arbitrators. Generally, the award must be based upon the law as determined by the parties in their agreement. This failing, the arbitrator must apply the law he considers proper in accordance with the rules of conflict of laws. In both cases, the arbitrator will have to take account of the terms of the contract and the usage of the specific trade. If, during any arbitration proceeding, a compromise is reached by the parties, that compromise may be recorded as an award by the arbitrator.

Appeals to the courts from the award cannot be excluded by agreement of the parties, since the fairness of the arbitration process as a quasi-judicial procedure has to be maintained. Any court control is, however, confined to specific matters, usually enumerated in the arbitration statutes, such as misconduct of the arbitrator in denying a party the full presentation of its claim or not granting a postponement of the hearing for good cause. A review of the award by a court generally will not deal with the arbitrator's decisions as to facts or with his application of the law. The competence of the courts is restricted in order not to make the arbitration process the beginning of litigation instead of its end. Recognition of an award and its enforcement will be denied when it appears to be contrary to public policy, as might be the case, for example, in cases involving trusts (monopolies), industrial property rights, or violation of foreign-currency restrictions. An arbitration award has the authority of a court decision and may be enforced by summary court action according to the procedural law of the country in which execution is being sought.

International commercial arbitration.

International commercial arbitration between traders of different countries has long been recognized by the business community and the legal profession as a suitable means of settling trade controversies out of court. The procedure in international commercial arbitration is basically the same as in domestic arbitration. In order to establish more uniformity in procedure and to make access to arbitration facilities

more easily available, the United Nations economic commissions in 1966 published new rules applying to international arbitration. Those for Europe are contained in the "Arbitration Rules of the United Nations Economic Commission for Europe" and for Asia and the Far East in the "Economic Commission for Asia and the Far East Rules for International Commercial Arbitration."

The development of international commercial arbitration has been furthered by uniform arbitration legislation prepared by the United Nations Conference on International Commercial Arbitration in 1958 and by the Council of Europe and the Inter-American Juridical Committee of the Organization of American States. One particularly difficult problem of international commercial arbitration is the enforcement of awards in a country other than the one in which they were rendered. Statutory municipal laws do not usually contain provisions for the enforcement of foreign awards, and parties are faced with uncertainty about the law and practice of enforcement procedure in a country other than their own.

The aforementioned international agreements, to which most of the trading nations of the world adhere, facilitate the enforcement of foreign awards to the extent that no further action is necessary in the country in which the award was rendered: the opposing debtor must establish that the award has been set aside or that its effects have been suspended by a competent authority, thus shifting the burden of proof of the nonbinding character of the award to the losing party.

Further development of international commercial arbitration is encouraged by the United Nations Commission on International Trade Law, which aims at promoting the harmonization and unification of laws in the field of international commercial arbitration.

LABOUR ARBITRATION

Labour arbitration--the reference of disputes between management and labour unions to an impartial third party for a final resolution--is usually the last step under a collective-bargaining agreement after all other measures to achieve a settlement have been exhausted.

Labour arbitration is not, as is commercial arbitration, an auxiliary avenue of justice and thereby a substitute for ordinary court procedure but a technique also for settling or avoiding strikes.

Two major aspects of labour arbitration are usually distinguished: arbitration of rights and arbitration of interests. Arbitration of rights refers to the arbitration of an existing labour contract when a dispute over the application of that contract arises between labour and management. Arbitration of interests refers to arbitration between labour and management during the negotiation of a new labour contract.

Arbitration of rights.

Arbitration of rights under the terms of a collective-bargaining agreement is employed in the United States far more than in most other countries. Outside the United States, labour courts, industrial courts, or conciliation and arbitration commissions perform the function of arbitrating rights. These bodies are usually appointed by the government, and recourse to them is frequently compulsory.

More than 90 percent of the collective-bargaining agreements in the United States provide for arbitration as a last step in the grievance procedure. Employees, for example, through their union, may present for arbitration complaints concerning such matters as discipline, discharge, and violations of working conditions. Other issues frequently submitted to arbitration customarily concern premium payments and incentive rates, overtime and vacations, Christmas bonuses, seniority rights, and fringe benefits, such as pension and welfare plans.

The arbitrator's decision must be based on the collective-bargaining agreement, which provides for the application of an existing contract to the grievance presented. The question of whether the various steps in the grievance procedure have been complied with before the initiation of the arbitration is usually left to the determination of the arbitrator and not of a court. The question, however, of whether the disputed issue is covered by the collective-bargaining agreement has to be determined by a court and not by the arbitrator. This authority of the courts was upheld by the Supreme Court of the United States in 1960.

The choice of arbitrator is made either by naming him in the agreement or, more often, by leaving the choice open until a dispute has arisen. Frequently, only a single arbitrator is appointed--usually an expert in the field of industrial relations. Alternatively, tripartite arbitration boards are established, both parties appointing their own arbitrator, who acts somewhat as advocate. A neutral chairman is selected either by the parties or by the two party-appointed arbitrators.

A further technique of arbitration of rights is the appointment of a single permanent arbitrator, or "umpire," to resolve disputes for the duration of the collective-bargaining agreement. The umpire will be intimately acquainted with the various economic, financial, and other aspects of the particular industry and will be familiar with the relationship between management and union developed in the past. He sometimes follows precedents, especially those established by his predecessor. This permanent umpire system originated in the United States in the anthracite-coal industry at the beginning of the 20th century and has been employed in such other important industries as newspaper printing and clothing.

Labour arbitrators render binding decisions on the disputes submitted to them. They are not bound by strict rules of court procedure, especially as regards burden of proof and the presentation of evidence. As arbitrators, they have the power to subpoena persons and written evidence. Labour arbitrators tend to evaluate factual evidence rather freely and often reduce penalties imposed upon employees by the management for breach of the labour contract. Even minor questions, such as the use of company time by employees for wash-ups or coffee breaks, are submitted to arbitration, in order to establish precedents in the operation of the plant. Generally, however, arbitrators are not bound to follow previous decisions.

Decisions of labour arbitrators are seldom reviewed by the courts, as awards are usually fully complied with by both parties.

Arbitration of interests.

Arbitration of the terms of a new contract, referred to as arbitration of interests, may be instituted if management and the labour union are unable to agree on a new contract. In some industries, such as hotels, printing, transit, and utilities, such

disputes are submitted to arbitration. In most countries, however, management and union are seldom inclined to forgo resort to lockouts and strikes in an attempt to obtain favourable new contracts, and interest arbitration is thus seldom used.

Compulsory arbitration, directed by legislative fiat, has been a controversial issue in the settlement of industrial disputes. It has been favoured in disputes in the transportation industry, which may involve great public inconvenience, and in disputes in the public-utilities sector when an immediate danger to public health and safety might occur. Compulsory arbitration has been declared unconstitutional in some states of the United States. More recently, however, it has been adopted as a regular procedure for the settlement of disputes with municipal employees in some U.S. cities.

INTERNATIONAL ARBITRATION

Controversies between sovereign states that are not settled by diplomatic negotiation or conciliation are often referred, by agreement of both parties, to the decision of a third disinterested party, who arbitrates the dispute with binding force upon the disputant parties. Such arbitration between states has a long history: it was used between city-states in ancient Greece and also in the Middle Ages, when the pope often acted as the sole arbitrator.

Historical development.

The modern development of international arbitration can be traced to the Jay Treaty of 1794 between Great Britain and the United States, which established three arbitral commissions to settle questions and claims arising out of the American Revolution. In the 19th century many arbitral agreements were concluded by which arbitration tribunals were established ad hoc to deal with a specific case or to handle a great number of claims. Most significant was the "Alabama" claim arbitration under the Treaty of Washington (1871), by which the United States and Great Britain agreed to settle claims arising from the failure of Great Britain to maintain its neutrality during the U.S. Civil War.

Commissions consisting of members drawn from both disputant countries ("mixed arbitral commissions") were often used in the 19th century to settle pecuniary

claims for compensation of injuries to aliens for which justice could not be obtained in foreign courts. Such was the purpose of a convention of 1868 between the United States and Mexico, by which claims of citizens of each country arising from the Civil War were settled. Boundary disputes between states were also often settled by arbitration.

International arbitration was given a more permanent basis by the Hague Conference of 1899, which adopted a convention on the pacific settlement of international disputes, revised by a Conference of 1907. The convention stated that: International arbitration has for its object the settlement of disputes between States by judges of their own choice and on the basis of respect for law. Recourse to arbitration implies an engagement to submit in good faith to the award.

A Permanent Court of Arbitration, composed of a panel of jurists appointed by the member governments, from which the litigant governments may select the arbitrators, was established at The Hague in 1899.

Twenty cases were arbitrated between 1902 and 1932, but from that year until 1972 only five cases were dealt with. This was largely because the importance of the Permanent Court of Arbitration was lessened by the Permanent Court of Justice (established in 1922) and its successor, the International Court of Justice. More recently, in 1960, the court, which was originally devised for the settlement of disputes between states, has offered its services for the arbitration of controversies between states and individuals or corporations.

Such a dispute was arbitrated in 1970 between a British company and the government of the Democratic Republic of Sudan. The case concerned the repudiation of a contract for building houses in the irrigation zone of the Khashm al-Qirbah Dam in The Sudan.

Arbitration provisions of international treaties.

There are several multilateral treaties that provide for the pacific settlement of international disputes by arbitration, including the Geneva General Act for the Settlement of Disputes of 1928, adopted by the League of Nations and reactivated by the General Assembly of the United Nations in 1949, which provides for the

settlement of various disputes, after unsuccessful efforts at conciliation, by an arbitral tribunal of five members. Other such treaties include the General Treaty of Inter-American Arbitration, signed in Washington in 1929, and the American Treaty on Pacific Settlement of Disputes, signed in Bogotá in 1948. More recently the Council of Europe adopted the European Convention for the Peaceful Settlement of Disputes (1957).

Arbitration is also mentioned as a proper method of settling disputes between countries in the Charter of the United Nations, as it was in the Covenant of the League of Nations.

The International Law Commission of the United Nations submitted to the General Assembly in 1955 a Convention on Arbitral Procedure. Its model rules would not become binding on any member-state of the United Nations unless they were accepted by a state in an arbitration treaty or in a special arbitral agreement. The model rules, however, have not yet been adopted in any arbitration arrangement between disputant governments, although in 1958 the General Assembly recommended the model rules for use by member-states when appropriate. It seems clear that states prefer flexibility in the resolution of their disputes by arranging the rules and proceedings of an arbitration according to circumstances.

Great impediments, indeed, still exist in the acceptance of international arbitration, especially in cases in which disputes between governments and foreign private parties are involved. In such cases the state will often insist that its own local remedies--administrative and court proceedings--be exhausted. Generally, the government of the national who advances a claim against a foreign government will require evidence that the injured party has pursued all remedies in the foreign country before it presses a claim for international negotiation and adjudication, if, indeed, it decides to take up the case at all. Contracting parties may agree in their contract that they need not exhaust local remedies before resorting to arbitration, and one 1965 instrument, the Convention on the Settlement of Investment Disputes, states:

Consent of the parties to arbitration under this Convention shall, unless otherwise stated, be deemed consent to such arbitration to the exclusion of any other remedy. A Contracting State may require the exhaustion of local administrative or judicial remedies as a condition of its consent to arbitration under this Convention.

The arbitration agreement in a general multilateral treaty, a bilateral convention, or in a specific contractual arrangement between two states often does not deal with the selection of the arbitrators and the appointment of an umpire, the procedure to be followed in the arbitration, the subject matter of the dispute, the specific issues to be submitted, the presentation of evidence, the place of the hearings, the law to be applied by the arbitrators, and the time when the award has to be rendered. These questions are usually dealt with in an agreement between the parties to the dispute known as compromis. If the compromis fails in some particular--to define the applicable law, for example--it is generally understood that the arbitrator shall apply international law.

An award rendered by an arbitral tribunal is customarily complied with by states: it is, in fact, invariably the case that unless a state is prepared to comply with an adverse decision, it will not submit the dispute to arbitration. The difficulties in the use of international arbitration thus consist less in the enforcement of arbitral awards than in persuading states involved in disputes to submit them to a third party, an arbitrator, or an arbitration tribunal.

ИНСТИТУТ ЭКОЛОГИИ, МЕДИЦИНЫ И ФИЗИЧЕСКОЙ КУЛЬТУРЫ

ECOLOGY

Long unfamiliar to the public, and relegated to a second-class status by many in the world of science, ecology emerged in the late 20th century as one of the most popular and most important aspects of biology. It has become painfully evident that the most pressing problems in the affairs of men - expanding populations, food scarcities, environmental pollution, and all the attendant sociological and political problems - are to a great degree ecological.

The word ecology was coined by a German zoologist. Ernst Haeckel, who applied the term oekologie to the "relation of the animal both to its organic as well as its inorganic environment." The word comes from the Greek oikos, meaning "household, home, or place to live." Thus ecology deals with the organism and its environment. The word environment includes both other organisms and physical surroundings. It involves relationships between individuals within a population and between individuals of different populations. These interactions between individuals, between populations, and between organisms and their environment form ecological systems, or ecosystems. Ecology has been defined variously as "the study of the interrelationships of organisms with their environment and each other," as "the economy of nature," and as "the biology of ecosystems."

Historical background.

Ecology had no firm beginnings. It evolved from the natural history of the Greeks, particularly Theophrastus, a friend and associate of Aristotle. He first described the interrelationships between organisms and between organisms and their nonliving environment. Later foundations for modern ecology were laid in the early work of plant and animal physiologists.

In the early and mid-1900s two groups of botanists, one in Europe and the other in America, studied plant communities from two different points of view. The European botanists concerned themselves with the study of the composition,

structure, and distribution of plant communities. The American botanists studied the development of plant communities, or succession. Both plant and animal ecology developed separately until American biologists emphasized the interrelation of both plant and animal communities as a biotic whole.

During the same period interest in population dynamics developed. The study of population dynamics received special impetus in the early 19th century, after Thomas Malthus called attention to the conflict between expanding populations and the capability of the earth to supply food. R.Pearl (1920), A.J. Lotka (1925), and V. Volterra (1926) developed mathematical foundations for the study of populations, and these studies led to experiments on the interaction of predators and prey, competitive relationships between species, and the regulation of populations. Investigations of the influence of behaviour on populations was stimulated by the recognition in 1920 of territoriality in nesting birds. Concepts of instinctive and aggressive behaviour were developed by K. Lorenz and N. Tinbergen, and the role of social behaviour in the regulation of populations was explored by V.C. Wynne-Edwards.

While some ecologists were studying the dynamics of communities and populations, others were concerned with energy-budgets. In 1920, August Thienemann, a German freshwater biologist, introduced the concept of trophic, or feeding, levels, by which the energy of food is transferred through a series of organisms, from green plants (the producers) up to several levels of animals (the consumers). An English animal ecologist, C.E. Elton (1927), further developed this approach with the concept of ecological niches and pyramids of numbers. Two American freshwater biologists, E. Birge and C. Juday, in the 1930s, in measuring the energy budgets of lakes, developed the idea of primary production, i.e., the rate at which food energy is generated, or fixed, by photosynthesis. Modern ecology came of age in 1942 with the development, by R.L. Lindeman of the United States, of the trophic-dynamic concept of ecology, which details the flow of energy through the ecosystem. Quantified field studies of energy flow through ecosystems were further developed by Eugene and Howard Odum of the United States; similar

early work on the cycling of nutrients was done by J.D. Ovington of England and Australia.

The study of both energy flow and nutrient cycling was stimulated by the development of new techniques - radioisotopes, microcalorimetry, computer science, and applied mathematics - that enabled ecologists to label, trace, and measure the movement of particular nutrients and energy through the ecosystems. These modern methods encouraged a new stage in the development of ecology - systems ecology, which is concerned with the structure and function of ecosystems.

Until the late 20th century ecology lacked a strong conceptual base. Modern ecology, however, is now focussed on the concept of the ecosystem, a functional unit consisting of interacting organisms and all aspects of the environment in any specific area. It contains both the nonliving (abiotic) and living (biotic) components through which nutrients are cycled and energy flows. To accomplish this cycling and flow, ecosystems must possess a number of structured interrelationships between soil, water, and nutrients, on the one hand, and producers, consumers, and decomposers on the other.

Ecosystems function by maintaining a flow of energy and a cycling of materials through a series of steps of eating and being eaten, of utilization and conversion, called the food chain. Ecosystems tend toward maturity, or stability, and in doing so they pass from a less complex to a more complex state. This directional change is called succession. Whenever an ecosystem is used, and that exploitation is maintained - as when a pond is kept clear of encroaching plants or a woodland is grazed by domestic cattle - the maturity of the ecosystem is effectively postponed.

The major functional unit of the ecosystem is the population. It occupies a certain functional niche, related to its role in energy flow and nutrient cycling. Both theenvironment and the amount of energy fixation in any given ecosystem are limited. When a population reaches the limits imposed by the ecosystem, its numbers must stabilize or, failing this, decline from disease, starvation, strife, low reproduction, or other behavioral and physiological reactions.

Changes and fluctuations in the environment represent selective pressure upon the population to which it must adjust. The ecosystem has historical aspects: the present is related to the past and the future to the present. Thus the ecosystem is the one concept that unifies plant and animal ecology, population dynamics, behaviour, and evolution.

Areas of study.

Of necessity, ecology is a multidisciplinary science. It involves plant and animal biology, taxonomy, physiology, genetics, behaviour, meteorology, pedology, geology, sociology, anthropology, physics, chemistry, mathematics, and electronics. Often it is difficult to draw a sharp line between ecology and any of these, for all impinge on it. The same situation exists also within ecology. In understanding the interactions between the organism and the environment or between organisms, it is often difficult to separate behaviour from population dynamics, behaviour from physiology, adaptation from evolution and genetics, animal ecology from plant ecology.

Ecology developed along two lines: the study of plants and the study of animals. Plant ecology concerns the relationships of plants to other plants and their environment. The approach is largely descriptive of the vegetational and floristic composition of an area and usually ignores the influence of animals on the plants. Animal ecology concerns the study of population dynamics, distribution, behaviour, and the interrelationships of animals and their environment. Because animals depend upon plants for food and shelter, animal ecology cannot be fully understood without a considerable background of plant ecology. This is particularly true in applied areas of ecology - wildlife and range management.

Both plant and animal ecology may be approached as the study of the interrelations of an individual organism with its environment, called autecology, or as the study of groups of organisms, called synecology.

Autecology, in many ways the classical study of ecology, is experimental and inductive. Because it is usually concerned with the relationship of an organism to

one or more variables such as humidity, light, salinity, or nutrient levels, it is easily quantified and lends itself to experimental design both in the field and the laboratory. It has borrowed techniques from chemistry, physics, and physiology. Autecology has contributed at least two important concepts: the constancy of interaction between an organism and its environment, and the genetic adaptability of local populations to local environmental conditions.

Synecology, on the other hand, is philosophical and deductive. It is largely descriptive and not easily quantified and contains a bewildering array of terminology. Only recently, since the advent of the electronic and atomic ages, has synecology developed the tools to study complex systems and enter an experimental phase. Important concepts developed by synecology are those concerned with nutrient cycling, energy budgets, and ecosystem development. Synecology has strong ties with pedology, geology, meteorology, and cultural anthropology.

Synecology may be subdivided according to environmental types, as terrestrial or aquatic. Terrestrial ecology, which may be further subdivided into forest, grassland, arctic, and desert ecology, concerns such aspects of terrestrial ecosystems as microclimate, soil chemistry, soil fauna, hydrologic cycles, ecogenetics, and productivity. Terrestrial ecosystems are more influenced by organisms and subject to much wider environmental fluctuations than are aquatic ecosystems. Aquatic ecosystems are affected more by the condition of the water and resist such environmental variables as temperature. Because the physical environment is so important in controlling aquatic ecosystems, considerable attention is paid to the chemical and physical characteristics of the ecosystem, such as the currents and the chemical composition of the water. By convention, aquatic ecology, called limnology, is limited to freshwater stream ecology and lake ecology. The former concerns life in flowing waters; the latter, life in relatively still water. Marine ecology deals with life in the open sea and in estuaries.

Other ecological approaches concern specialized areas. The study of the geographic distribution of plants and animals is ecological plant and animal

geography. The study of population growth, mortality, natality, competition, and predator-prey relations is population ecology. The study of the genetics and ecology of local races and distinct species is ecological genetics. The study of the behavioral responses of animals to their environment, and of social interactions as they affect population dynamics, is behavioral ecology. Investigations of interactions between the physical environment and the organism fall under ecoclimatology and physiological ecology. The study of groups of organisms is community ecology (though it is difficult to separate it from studies of bioenergetics, biogeochemical cycles, and trophic-dynamic aspects of the community or ecosystem ecology). That part of ecosystem ecology concerned with the analysis and understanding of the structure and function of ecosystems by the use of applied mathematics, mathematical models, and computer programs is systems ecology. Systems ecology, concentrating on input and output analysis, has stimulated the rapid development of applied ecology, concerned with the application of ecological principles to the management of natural resources, agricultural production, and problems of environmental pollution.

Methods in ecology.

Because ecologists work with living systems possessing numerous variables, the techniques used by physicists and chemists, mathematicians and engineers, require modification; they are not easily applied nor are the results as precise as those obtained in other sciences. It is relatively simple, for example, for a physicist to measure gain and loss of heat from metals or other inanimate objects, which possess certain constants of conductivity, expansion, surface features, and the like. To determine the heat exchange between an animal and its environment, however, a physiological ecologist is confronted with an array of almost unquantifiable variables and has the formidable task of both gathering the numerous data and analyzing them.

Ecological measurements probably never will be as precise or as subject to the same ease of analysis as measurements in physics, chemistry, or certain quantifiable areas of biology.

In spite of these problems, various aspects of the environment can be determined by physical and chemical means, ranging from simple chemical identifications and physical measurements to the use of sophisticated mechanical apparatus. The development of biostatistics and proper experimental design, and the improvements in methods of sampling, permit a quantified statistical approach to the study of ecology. Because of the extreme difficulties of controlling environmental variables in the field, studies involving the use of experimental design are largely confined to the laboratory and to controlled field experiments designed to test the effects of only one variable or several variables. The use of statistical procedures, and the application of computer science to mathematical models based on data obtained from the field, are providing new insights into population interactions and ecosystem function. Mathematical programming is becoming increasingly important in applied ecology, especially in the management of natural resources and agricultural problems having an ecological basis.

Controlled environmental chambers enable experimenters to maintain plants and animals under known conditions of light, temperature, humidity, and daylength so that the effects of each variable (or combination of variables) on the organism can be studied. Biotelemetry and other electronic tracking equipment, products of the space age, permit the rapid and nondestructive sampling of plant and animal populations. Such tools enable ecologists to follow from a distance the movements and behaviour of a free-ranging animal by radio signals beamed from a sender attached to the organism.

Radioisotopes are used for tracing the pathways of nutrients through ecosystems, for determining the time and extent of transfer of energy and nutrients through the different components of the ecosystem, and for the determination of food chains. The use of laboratory microcosms - aquatic and soil micro-ecosystems, consisting of biotic and nonbiotic material from natural ecosystems, held under conditions

similar to those found in the field—are useful in determining rates of nutrient cycling, ecosystem development, and other functional aspects of ecosystems. Microcosms enable the ecologist to duplicate experiments and to perform experimental manipulation on them.

ZOOLOGY

Zoology, the study of animals, includes both the inquiry into individual animals and their constituent parts, even to the molecular level, and the inquiry into animal populations, entire faunas, and the relationships of animals to each other, to plants, and to the nonliving environment. Though this wide range of studies results in some isolation of specialties within zoology, the conceptual integration in the contemporary study of living things that has occurred in recent years emphasizes the structural and functional unity of life rather than its diversity.

Historical background.

Prehistoric man's survival as a hunter defined his relation to other animals, which were a source of food and danger. As man's cultural heritage developed, animals were variously incorporated into man's folklore and philosophical awareness as fellow living creatures. Domestication of animals forced man to take a systematic and measured view of animal life, especially after urbanization necessitated a constant and large supply of animal products.

Study of animal life by the ancient Greeks became more rational, if not yet scientific, in the modern sense, after the cause of disease - until then thought to be demons - was postulated by Hippocrates to result from a lack of harmonious functioning of body parts. The systematic study of animals was encouraged by Aristotle's extensive descriptions of living things, his work reflecting the Greek concept of order in nature and attributing to nature an idealized rigidity.

In Roman times Pliny brought together in 37 volumes a treatise, Historia naturalis, that was an encyclopaedic compilation of both myth and fact regarding celestial bodies, geography, animals and plants, metals, and stone. Volumes VII to XI concern zoology; volume VIII, which deals with the land animals, begins with the largest one, the elephant. Although Pliny's approach was naïve, his scholarly effort had a profound and lasting influence as an authoritative work.

Zoology continued in the Aristotelian tradition for many centuries in the Mediterranean region and by the Middle Ages, in Europe, it had accumulated considerable folklore, superstition, and moral symbolisms, which were added to otherwise objective information about animals.

Gradually, much of this misinformation was sifted out: naturalists became more critical as they compared directly observed animal life in Europe with that described in ancient texts. The use of the printing press in the 15th century made possible an accurate transmission of information. Moreover, mechanistic views of life processes (i.e., that physical processes depending on cause and effect can apply to animate forms) provided a hopeful method for analyzing animal functions; for example, the mechanics of hydraulic systems were part of William Harvey's argument for the circulation of the blood - although Harvey remained thoroughly Aristotelian in outlook. In the 18th century, zoology passed through reforms provided by both the system of nomenclature of Carolus Linnaeus and the comprehensive works on natural history by Georges-Louis Leclerc de Buffon; to these were added the contributions to comparative anatomy by Georges Cuvier in the early 19th century.

Physiological functions, such as digestion, excretion, and respiration, were easily observed in many animals, though they were not as critically analyzed as was blood circulation.

Following the introduction of the word cell in the 17th century and microscopic observation of these structures throughout the 18th century, the cell was incisively defined as the common structural unit of living things in 1839 by two Germans: Matthias Schleiden and Theodor Schwann. In the meanwhile, as the science of chemistry developed, it was inevitably extended to an analysis of animate systems. In the middle of the 18th century the French physicist René Antoine Ferchault de Réaumer demonstrated that the fermenting action of stomach juices is a chemical process. And in the mid-19th century the French physician and physiologist Claude Bernard drew upon both the cell theory and knowledge of chemistry to develop the concept of the stability of the internal bodily environment, now called homeostasis.

The cell concept influenced many biological disciplines, including that of embryology, in which cells are important in determining the way in which a fertilized egg develops into a new organism. The unfolding of these events - called epigenesis by Harvey - was described by various workers, notably the German-trained comparative embryologist Karl von Baer, who was the first to observe a mammalian egg within an ovary. Another German-trained embryologist, Christian Heinrich Pander, introduced in 1817 the concept of germ, or primordial, tissue layers into embryology.

In the latter part of the 19th century, improved microscopy and better staining techniques using aniline dyes, such as hematoxylin, provided further impetus to the study of internal cellular structure.

By this time Darwin had made necessary a complete revision of man's view of nature with his theory that biological changes in species occur through the process of natural selection. The theory of evolution - that organisms are continuously evolving into highly adapted forms - required the rejection of the static view that all species are especially created and upset the Linnaean concept of species types. Darwin recognized that the principles of heredity must be known to understand how evolution works; but, even though the concept of hereditary factors had by then been formulated by Mendel, Darwin never heard of his work, which was essentially lost until its rediscovery in 1900.

Genetics has developed in the 20th century and now is essential to many diverse biological disciplines. The discovery of the gene as a controlling hereditary factor for all forms of life has been a major accomplishment of modern biology. There has also emerged clearer understanding of the interaction of organisms with their environment. Such ecological studies help not only to show the interdependence of the three great groups of organisms - plants, as producers; animals, as consumers; and fungi and many bacteria, as decomposers - but they also provide information essential to man's control of the environment and, ultimately, to his survival on Earth. Closely related to this study of ecology are inquiries into animal behaviour, or ethology. Such studies are often cross disciplinary in that ecology, physiology,

genetics, development, and evolution are combined as man attempts to understand why an organism behaves as it does. This approach now receives substantial attention because it seems to provide useful insight into man's biological heritage - that is, the historical origin of man from nonhuman forms.

The emergence of animal biology has had two particular effects on classical zoology. First, and somewhat paradoxically, there has been a reduced emphasis on zoology as a distinct subject of scientific study; for example, workers think of themselves as geneticists, ecologists, or physiologists who study animal rather than plant material. They often choose a problem congenial to their intellectual tastes, regarding the organism used as important only to the extent that it provides favourable experimental material. Current emphasis is, therefore, slanted toward the solution of general biological problems; contemporary zoology thus is to a great extent the sum total of that work done by biologists pursuing research on animal material.

Second, there is an increasing emphasis on a conceptual approach to the life sciences. This has resulted from the concepts that emerged in the late 19th and early 20th centuries: the cell theory; natural selection and evolution; the constancy of the internal environment; the basic similarity of genetic material in all living organisms; and the flow of matter and energy through ecosystems. The lives of microbes, plants, and animals now are approached using theoretical models as guides rather than by following the often restricted empiricism of earlier times. This is particularly true in molecular studies, in which the integration of biology with chemistry allows the techniques and quantitative emphases of the physical sciences to be used effectively to analyze living systems.

Areas of study.

Although it is still useful to recognize many disciplines in animal biology - e.g., anatomy or morphology; biochemistry and molecular biology; cell biology; developmental studies (embryology); ecology; ethology; evolution; genetics;

physiology; and systematics - the research frontiers occur as often at the interfaces of two or more of these areas as within any given one.

Anatomy or morphology.

Descriptions of external form and internal organization are among the earliest records available regarding the systematic study of animals. Aristotle was an indefatigable collector and dissector of animals. He found differing degrees of structural complexity, which he described with regard to ways of living, habits, and body parts. Although Aristotle had no formal system of classification, it is apparent that he viewed animals as arranged from the simplest to the most complex in an ascending series. Since man was even more complex than animals and, moreover, possessed a rational faculty, he therefore occupied the highest position and a special category. This hierarchical perception of the animate world proved to be useful in every century to the present, except that in the modern view there is no such "scale of nature," and there is change in time by evolution from the simple to the complex.

After the time of Aristotle, Mediterranean science was centred at Alexandria, where the study of anatomy, particularly the central nervous system, flourished and, in fact, first became recognized as a discipline. Galen studied anatomy at Alexandria in the 2nd century and later dissected many animals. Much later, the contributions of the Renaissance anatomist Andreas Vesalius, though made in the context of medicine, as were those of Galen, stimulated to a great extent the rise of comparative anatomy. During the latter part of the 15th century and throughout the 16th century, there was a strong tradition in anatomy; important similarities were observed in the anatomy of different animals, and many illustrated books were published to record these observations.

But anatomy remained a purely descriptive science until the advent of functional considerations in which the correlation between structure and function was consciously investigated; as by French biologists Buffon and Cuvier. Cuvier cogently argued that a trained naturalist could deduce from one suitably chosen

part of an animal's body the complete set of adaptations that characterized the organism. Because it was obvious that organisms with similar parts pursue similar habits, they were placed together in a system of classification. Cuvier pursued this viewpoint, which he called the theory of correlations, in a somewhat dogmatic manner and placed himself in opposition to the romantic natural philosophers, such as the German intellectual Johann Wolfgang von Goethe, who saw a tendency to ideal types in animal form. The tension between these schools of thought - adaptation as the consequence of necessary bodily functions and adaptation as an expression of a perfecting principle in nature - runs as a leitmotiv through much of biology, with overtones extending into the early 20th century.

The twin concepts of homology (similarity of origin) and analogy (similarity of appearance), in relation to structure, are the creation of the 19th-century British anatomist Richard Owen. Although they antedate the Darwinian view of evolution, the anatomical data on which they were based became, largely as a result of the work of the German comparative anatomist Carl Gegenbaur, important evidence in favour of evolutionary change, despite Owen's steady unwillingness to accept the view of diversification of life from a common origin.

In summary, anatomy moved from a purely descriptive phase as an adjunct to classificatory studies, into a partnership with studies of function and became, in the 19th century, a major contributor to the concept of evolution.

Taxonomy or systematics.

Not until the work of Carolus Linnaeus did the variety of life receive a widely accepted systematic treatment. Linnaeus strove for a "natural method of arrangement," one that is now recognizable as an intuitive grasp of homologous relationships, reflecting evolutionary descent from a common ancestor; however, the natural method of arrangement sought by Linnaeus was more akin to the tenets of idealized morphology because he wanted to define a "type" form as epitomizing a species.

It was in the nomenclatorial aspect of classification that Linnaeus created a revolutionary advance with the introduction of a Latin binomialsystem: each species received a Latin name, which was not influenced by local names and which invoked the authority of Latin as a language common to the learned people of that day. The Latin name has two parts. The first word in the Latin name for the dog, Canis familiaris, for example, indicates the larger category, or genus, to which dogs belong; the second word is the name of the species within the genus. In addition to species and genera, Linnaeus also recognized other classificatory groups, or taxa (singular taxon), which are still used; namely, order, class, and kingdom, to which have been added family (between genus and order) and phylum (between class and kingdom). Each of these can be divided further by the appropriate prefix of sub- or super-, as in subfamily or superclass. Linnaeus' great work, the Systema naturae, went through 12 editions during his lifetime; the 13th, and final, edition appeared posthumously. Although his treatment of the diversity of living things has been expanded in detail, revised in terms of taxonomic categories, and corrected in the light of continuing work - for example, Linnaeus treated whales as fish - it still sets the style and method, even to the use of Latin names, for contemporary nomenclatorial work.

Linnaeus sought a natural method of arrangement, but he actually defined types of species on the basis of idealized morphology. The greatest change from Linnaeus' outlook is reflected in the phrase "the new systematics," which was introduced in the 20th century and through which an explicit effort is made to have taxonomic schemes reflect evolutionary history. The basic unit of classification, the species, is also the basic unit of evolution - i.e., a population of actually or potentially interbreeding individuals. Such a population shares, through interbreeding, its genetic resources. In so doing, it creates the gene pool - its total genetic material - that determines the biological resources of the species and on which natural selection continuously acts. This approach has guided work on classifying animals away from somewhat arbitrary categorization of new species to that of recreating evolutionary history (phylogeny) and incorporating it in the system of

classification. Modern taxonomists or systematists, therefore, are among the foremost students of evolution.

Physiology.

The practical consequences of physiology have always been an unavoidable human concern, in both medicine and animal husbandry. Inevitably, from Hippocrates to the present, practical knowledge of human bodily function has accumulated along with that of domestic animals and plants. This knowledge has been expanded, especially since the early 1800s, by experimental work on animals in general, a study known as comparative physiology. The experimental dimension had wide applications following Harvey's demonstration of the circulation of blood. From then on, medical physiology developed rapidly; notable texts appeared, such as Albrecht von Haller's eight-volume work Elementa Physiologiae Corporis Humani (Elements of Human Physiology), which had a medical emphasis. Toward the end of the 18th century the influence of chemistry on physiology became pronounced through Antoine Lavoisier's brilliant analysis of respiration as a form of combustion. This French chemist not only determined that oxygen was consumed by living systems but also opened the way to further inquiry into the energetics of living systems. His studies further strengthened the mechanistic view, which holds that the same natural laws govern both the inanimate and the animate realms.

Physiological principles achieved new levels of sophistication and comprehensiveness with Bernard’s concept of constancy of the internal environment, the point being that only under certain constantly maintained conditions is there optimal bodily function. His rational and incisive insightswere augmented by concurrent developments in Germany, where Johannes Müller explored the comparative aspects of animal function and anatomy, and Justus von Liebig and Car1 Ludwig applied chemical and physical methods, respectively, to the solution of physiological problems. As a result, many useful techniques were advanced - e.g., means for precise measurement of muscular action and changes in blood pressure and means for defining the nature of body fluids.

By this time the organ systems - circulatory, digestive, endocrine, excretory, integumentary, muscular, nervous, reproductive, respiratory, and skeletal - had been defined, both anatomically and functionally, and research efforts were

focussed on understanding these systems in cellular and chemical terms, an emphasis that continues to the present and has resulted in specialties in cell physiology and physiological chemistry. General categories of research now deal with the transportation of materials across membranes; the metabolism of cells, including synthesis and breakdown of molecules; and the regulation of these processes.

Interest has also increased in the most complex of physiological systems, the nervous system. Much comparative work has been done by utilizing animals with structures especially amenable to various experimental techniques; for example, the large nerves in squids have been extensively studied in terms of the transmission of nerve impulses, and insect and crustacean eyes have yielded significant information on patterns of sensory inputs. Most of this work is closely associated with studies on animal orientation and behaviour. Although the contemporary physiologist often studies functional problems at the molecular and cellular levels, he is also aware of the need to integrate cellular studies into the many-faceted functions of the total organism.

Embryology, or developmental studies.

Embryonic growth and differentiation of parts have been major biological problems since ancient times. A 17th-century explanation of development assumed that the adult existed as a miniature - a homunculus - in the microscopic material that initiates the embryo. But in 1759 the German physician Caspar Friedrick Wolff firmly introduced into biology the interpretation that undifferentiated materials gradually become specialized, in an orderly way, into adult structures. Although this epigenetic process is now accepted as characterizing the general nature of development in both plants and animals, many questions remain to be solved. The French physician Marie François Xavier Bichat declared in 1801 that differentiating parts consist of various components called tissues; with the subsequent statement of the cell theory, tissues were resolved into their cellular

constituents. The idea of epigenetic change and the identification of structural components made possible a new interpretation of differentiation.

It was demonstrated that the egg gives rise to three essential germ layers out of which specialized organs, with their tissues, subsequently emerge. Then, following his own discovery of the mammalian ovum, von Baer in 1828 usefully applied this information when he surveyed the development of various members of the vertebrate groups. At this point, embryology, as it is now recognized, emerged as a distinct subject.

The concept of cellular organization had an effect on embryology that continues to the present day. In the 19th century, cellular mechanisms were considered essentially to be the basis for growth, differentiation, and morphogenesis, or molding of parts. The distribution of the newly formed cells of the rapidly dividing zygote (fertilized egg) was precisely followed to provide detailed accounts not only of the time and mode of germ layer formation but also of the contribution of these layers to the differentiation of tissues and organs. Such descriptive information provided the background for experimental work aimed at elucidating the role of chromosomes and other cellular constituents in differentiation. About 1895, before the formulation of the chromosomal theory of heredity, Theodor Boveri demonstrated that chromosomes show continuity from one cell generation to the next. In fact, biologists soon concluded that in all cells arising from a fertilized egg, half the chromosomes are of maternal and half of paternal origin.

The discovery of the constant transmission of the original chromosomal endowment to all cells of the body served to deepen the mystery surroundingthe factors that determine cellular differentiation.

The present view is that differential activity of genes is the basis for cellular and tissue differentiation; that is, although the cells of a multicellular body contain the same genetic information, different genes are active in different cells. The result is the formation of various gene products, which regulate the functional and structural differentiation of cells. The actual mechanism involved in the inactivation of certain genes and the activation of others, however, has not yet been

established. That cells can move extensively throughout the embryo and selectively adhere to other cells, thus starting tissue aggregations, also contributes to development as does the fate of cells - i.e., certain ones continue to multiply, others stop, and some die.

Research methods in embryology now exploit many experimental situations: both unicellular and multicellular forms; regeneration (replacement of lost parts) and normal development; and growth of tissues outside and inside the host. Hence, the processes of development can be studied with material other than embryos; and the study of embryology has become incorporated into the more inclusive subdiscipline of developmental biology.

Evolutionism.

Darwin was not the first to speculate that organisms can change from generation to generation and so evolve, but he was the first to propose a mechanism by which the changes are accumulated. He proposed that heritable variations occur in conjunction with a never-ending competition for survival and that the variations favouring survival are automatically preserved. In time, therefore, the continued accumulation of variations results in the emergence of new forms. Because the variations that are preserved relate to survival, the survivors are highly adapted to their environment. To this process Darwin gave the apt name natural selection.

Many of Darwin's predecessors, notably Jean-Baptiste Lamarck, were willing to accept the idea of species variation, even though to do so meant denying the doctrine of special creation and the static-type species of Linnaeus. But they argued that some idealized perfecting principle, expressed through the habits of an organism, was the basis of variation. The contrast between the romanticism of Lamarck and the objective analysis of Darwin clearly reveals the type of revolution provoked by the concept of natural selection. Although mechanistic explanations had long been available to biologists - forming, for example, part of Harvey's explanation of blood circulation - they did not pervade the total structure of biological thinking until the advent of Darwinism.

There were two immediate consequences of Darwin's viewpoints. One has involved a reappraisal of all subject areas of biology; reinterpretations of morphology and embryology are good examples. The comparative anatomy of the British anatomist Owen became a cornerstone of the evidence forevolution, and German anatomists provided the basis for the comment that evolutionary thinking was born in England but gained its home in Germany. The reinterpretation of morphology carried over into the study of fossil forms, as paleontologists sought and found evidence of gradual change in their study of fossils. But some workers, although accepting evolution in principle, could not easily interpret the changes in terms of natural selection. The German paleontologist Otto Schindewolf, for example, found in shelled mollusks called ammonites evidence of progressive complexity and subsequent simplification of forms. The American paleontologist George Gaylord Simpson, however, has been a consistent interpreter of vertebrate fossils by Darwinian selection. Embryology was seen in an evolutionary light when the German zoologist Ernst Haeckel proposed that the epigenetic sequence of embryonic development (ontogeny) repeated its evolutionary history (phylogeny). Thus, the presence of gill clefts in the mammalian embryo and also in less highly evolved vertebrates can be understood as a remnant of a common ancestor.

The other consequence of Darwinism - to make more explicit the origin and nature of heritable variations and the action of natural selection on them - depended on the emergence of the following: genetics and the elucidation of the rules of Mendelian inheritance; the concept of the gene as the unit of inheritance; and the nature of gene mutation. The development of these ideas provided the basis for the genetics of natural populations.

The subject of population genetics began with the Mendelian laws of inheritance and now takes into account selection, mutation, migration (movement into and out of a given population), breeding patterns, and population size. These factors affect the genetic makeup of a group of organisms that either interbreed or have the potential to do so; i.e., a species. Accurate appraisal of these factors allows precise predictions regarding the content of a given gene pool over significant periods of

evolutionary time. From work involving population genetics has come the realization, eloquently documented by two contemporary American evolutionists, Theodosius Dobzhansky and Ernst Mayer, that the species is the basic unit of evolution. The process of speciation occurs as a gene pool breaks up to form isolated gene pools. When selection pressures similar to those of the original gene pool persist in the new gene pools, similar functions and the similar structures on which they depend also persist. When selection pressures differ, however, differences arise. Thus, the process of speciation through natural selection preserves the evolutionary history of a species. The record may be discerned not only in the gross, or macroscopic, anatomy of organisms but also in their cellular structure and molecular organization. Significant work now is carried out, for example, on the homologies of the nucleic acids and proteins of different species.

Genetics.

The problem of heredity had been the subject of careful study before its definitive analysis by Mendel. As with Darwin's predecessors, those of Mendel tended to idealize and interpret all inherited traits as being transmitted through the blood or as determined by various "humors" or other vague entities in animal organisms. When studying plants, Mendel was able to free himself of anthropomorphic and holistic explanations. By studying seven carefully defined pairs of characteristics - e.g., tall and short plants; red and white flowers, etc. - as they were transmitted through as many as three successive generations, he was able to establish patterns of inheritance that apply to all sexually reproducing forms. Darwin, who was searching for an explanation of inheritance, apparently never saw Mendel's work, which was published in 1866 in the obscure journal of his local natural history society; it was simultaneously rediscovered in 1900 by three different European geneticists.

Further progress in genetics was made early in the 20th century, when it was realized that heredity factors are found on chromosomes. The term gene was coined for these factors. Studies by the American geneticist Thomas Hunt Morgan

on the fruit fly (Drosophila), moved animal genetics to the forefront of genetic research. The work of Morgan and his students established such major concepts as the linear array of genes on chromosomes; the exchange of parts between chromosomes; and the interaction of genes in determining traits, including sexual differences. In 1927 one of Morgan's former students, Hermann Muller, used X-rays to induce the mutations (changes in genes) in the fruit fly, thereby opening the door to major studies on the nature of variation.

Meanwhile, other organisms were being used for genetic studies, most notably fungi and bacteria. The results of this work provided insights into animal genetics just as principles initially obtained from animal genetics provided insight into botanical and microbial forms. Work continues not only on the genetics of humans, domestic animals, and plants but also on the control of development through the orderly regulation of gene action in different cells and tissues.

Cellular and molecular biology.

Although the cell was recognized as the basic unit of life early in the 19th century, its most exciting period of inquiry has probably occurred since the 1940s. The new techniques developed since that time, notably the perfection of the electron microscope and the tools of biochemistry, have changed the cytological studies of the 19th and early 20th centuries from a largely descriptive inquiry, dependent on the light microscope, into a dynamic, molecularly oriented inquiry into fundamental life processes.

The so-called cell theory, which was enunciated about 1838, was never actually a theory. As Edmund Beecher Wilson, the noted American cytologist, stated in his great work, The Cell.

By force of habit we still continue to speak of the cell 'theory' but it is a theory only in name. In substance it is a comprehensive general statement of fact and as such stands today beside the evolution theory among the foundationst ones of modern biology.

More precisely, the cell doctrine was an inductive generalization based on the microscopial examination of certain plant and animal species.

Rudolf Virchow, a German medical officer specializing in cellular pathology, first expressed the fundamental dictum regarding cells in his phrase omnis cellula e cellula (all cells from cells). For cellular reproduction is the ultimate basis of the continuity of life; the cell is not only the basic structural unit of life but also the basic physiological and reproductive unit. All areas of biology were affected by the new perspective afforded by the principle of cellular organization. Especially in conjunction with embryology was the study of the cell most prominent in animal biology. The continuity of cellular generations by reproduction also had implications for genetics. It is little wonder, then, that the full title of Wilson's survey of cytology at the turn of the century was The Cell: Its Role in Development and Heredity.

The study of the cell nucleus, its chromosomes, and their behaviour served as the basis for understanding the regular distribution of genetic material during both sexual and asexual reproduction. This orderly behaviour of the nucleus made it appear to dominate the life of the cell, for by contrast the components of the rest of the cell appeared to be randomly distributed.

The biochemical study of life had helped in the characterization of the major molecules of living systems - proteins, nucleic acids, fats, and carbohydrates - and in the understanding of metabolic processes. That nucleic acids are a distinctive feature of the nucleus was recognized after their discovery by the Swiss biochemist Johann Friedrich Miescher in 1869. In 1944 a group of American bacteriologists, led by Oswald T. Avery, published work on the causative agent of pneumonia in mice (a bacterium) that culminated in the demonstration that deoxyribonucleic acid (DNA) is the chemical basis of heredity. Discrete segments of DNA correspond to genes, or Mendel's hereditary factors. Proteins were discovered to be especially important for their role in determining cell structure and in controlling chemical reactions.

The advent of techniques for isolating and characterizing proteins and nucleic acids now allows a molecular approach to essentially all biological problems - from the appearance of new gene products in normal development or under pathological conditions to a monitoring of changes in and between nerve cells during the transmission of nerve impulses.

Ecology.

The harmony that Linnaeus found in nature, which redounded to the glory and wisdom of a Judaeo-Christian god, was the 18th-century counterpart of the balanced interaction now studied by ecologists. Linnaeus recognized that plants are adapted to the regions in which they grow, that insects play a role in flower pollination, and that certain birds prey on insects and are in turn eaten by other birds. This realization implies, in contemporary terms, the flow of matter and energy in a definable direction through any natural assemblage of plants, animals, and microorganisms.

Such an assemblage, termed an ecosystem, starts with the plants, which are designated as producers because they maintain and reproduce themselves at the expense of energy from sunlight and inorganic materials taken from the nonliving environment around them (earth, air, and water).Animals are called consumers because they ingest plant material or other animals that feed on plants, using the energy stored in this food to sustain themselves. Lastly, the organisms known as decomposers, mostly fungi and bacteria, break down plant and animal material and return it to the environment in a form that can be used again by plants in a constantly renewed cycle.

The term ecology, first formulated by Haeckel in the latter part of the 19th century as "oecology" (from the Greek word for house, oikos), referred to the dwelling place of organisms in nature. In the 1890s various European and U.S. scientists laid the foundations for modern work through studies of natural ecosystems and the populations of organisms contained within them.

Animal ecology, the study of consumers and their interactions with the environment, is very complex; attempts to study it usually focus on one particular aspect. Some studies, for example, involve the challenge of the environment to individuals with special adaptations (e.g., water conservation in desert animals); others may involve the role of one species in its ecosystem or the ecosystem itself. Food-chain sequences have been determined for various ecosystems, and the efficiency of the transfer of energy and matter within them has been calculated so that their capacity is known; that is, productivity in terms of numbers of organisms or weight of living matter at a specific level in the food chain can be accurately determined.

In spite of advances in understanding animal ecology, this subject area of zoology does not yet have the major unifying theoretical principles found in genetics (gene theory) or evolution (natural selection).

Ethology.

The study of animal behaviour (ethology) is largely a 20th-century phenomenon and is exclusively a zoological discipline. Only animals have nervous systems, with their implications for perception, coordination, orientation, learning, and memory. Not until the end of the 19th century did animal behaviour become free from anthropocentric interests and assume an importance in its own right. The British behaviorist C. Lloyd Morgan was probably most influential with his emphasis on parsimonious explanations - i.e., that the explanation "which stands lower in the psychological scale" must be invoked first. This principle is exemplified in the American Herbert Spencer Jennings' pioneering work in 1906 on The Behavior of Lower Organisms.

The study of animal behaviour now includes many diverse topics, ranging from swimming patterns of protozoans to socialization and communication among the great apes. Many disparate hypotheses have been proposed in an attempt to explain the variety of behavioral patterns found in animals. They focus on the mechanisms that stimulate courtship in reproductive behaviour of such diverse groups as

spiders, crabs, and domestic fowl; and on whole life histories, starting from the special attachment of newly born ducks and goats to their actual mothers or to surrogate (substitute) mothers. The latter phenomenon, called imprinting, has been intensively studied by the Austrian ethologist Konrad Lorenz. Physiologically oriented behaviour now receives much attention; studies range from work on conditioned reflexes to the orientation of crustaceans and the location and communication of food among bees; such diversity of material is one measure of the somewhat diffuse but exciting current state of these studies.

General trends.

Zoology has become animal biology - that is, the life sciences display a new unity, one that is founded on the common basis of all life, on the gene pool-species organization of organisms, and on the obligatory interacting of the components of ecosystems. Even as regards the specialized features of animals - involving physiology, development, or behaviour - the current emphasis is on elucidating the broad biological principles that identify animals as one aspect of nature. Zoology has thus given up its exclusive emphasis on animals - an emphasis maintained from Aristotle's time well into the 19th century - in favour of a broader view of life. The successes in applying physical and chemical ideas and techniques to life processes have not only unified the life sciences but have also created bridges to other sciences in a way only dimly foreseen by earlier workers. The practical and theoretical consequences of this trend have just begun to be realized.

Methods in zoology.

Because the study of animals may be concentrated on widely different topics, such as ecosystems and their constituent populations, organisms, cells, and chemical reactions, specific techniques are needed for each kind of investigation. The emphasis on the molecular basis of genetics, development, physiology, behaviour, and ecology has placed increasing importance on those techniques involving cells and their many components. Microscopy, therefore, is a necessary technique in zoology, as are certain physicochemical methods for isolating and characterizing molecules. Computer technology also has a special role in the analysis of animal

life. These newer techniques are used in addition to the many classical ones - measurement and experimentation at the tissue, organ, organ system, and organismic levels.

Microscopy.

In addition to continuous improvements in the techniques of staining cells, so that their components can be seen clearly, the light used in microscopy can now be manipulated to make visible certain structures in living cells that are otherwise undetectable. The ability to observe living cells is an advantage of light microscopes over electron microscopes; the latter require the cells to be in an environment that kills them. The particular advantage of the electron microscope, however, is its great powers of magnification. Theoretically, it can resolve single atoms; in biology, however, magnifications of lesser magnitude are most useful in determining the nature of structures lying between whole cells and their constituent molecules.

Separation and purification techniques.

The characterization of components of cellular systems is necessary for biochemical studies. The specific molecular composition of cellular organelles, for example, affects their shape and density (mass per unit volume); as a result, cellular components settle at different rates (and thus can be separated) when they are spun in a centrifuge.

Other methods of purification rely on other physical properties. Molecules vary in their affinity for the positive or negative pole of an electrical field. Migration to or away from these poles, therefore, occurs at different rates for different molecules and allows their separation; the process is called electrophoresis. The separation of molecules by liquid solvents exploits the fact that the molecules differ in their solubility, and hence they migrate to various degrees as a solvent flows past them. This process, known as chromatography because of the colour used to identify the position of the migrating materials, yields samples of extraordinarily high purity.

Radioactive tracers.

Radioactive compounds are especially useful in biochemical studies involving metabolic pathways of synthesis and degradation. Radioactive compounds are incorporated into cells in the same way as their nonradioactive counterparts. These compounds provide information on the sites of specific metabolic activities within cells and insights into the fates of these compounds in both organisms and the ecosystem.

Computers.

Computers process information using their own general language, which is able to complete calculations as complex and diverse as statistical analyses and determinations of enzymatically controlled reaction rates. Computers with access to extensive data files can select information associated with a specific problem and display it to aid the researcher in formulating possible solutions. They help perform routine examinations such as scanning chromosome preparations in order to identify abnormalities in number or shape. Test organisms can be electronically monitored with computers, so that adjustments can be made during experiments; this procedure improves the quality of the data and allows experimental situations to be fully exploited. Computer simulation is important in analyzing complex problems; as many as 100 variables, for example, are involved in the management of salmon fisheries. Simulation makes possible the development of models that approach the complexities of conditions in nature, a procedure of great value in studying wildlife management and related ecological problems.

Applied zoology.

Animal-related industries produce food (meats and dairy products), hides, furs, wool, organic fertilizers, and miscellaneous chemical byproducts. There has been a dramatic increase in the productivity of animal husbandry since the 1870s, largely as a consequence of selective breeding and improved animal nutrition. The purpose of selective breeding is to develop livestock whose desirable traits have strong heritable components and can therefore be propagated. Heritable components are distinguished from environmental factors by determining the coefficient of

heritability, which is defined as the ratio of variance in a gene-controlled character to total variance.

Another aspect of food production is the control of pests. The serious side effects of some chemical pesticides make extremely important the development of effective and safe control mechanisms. Animal food resources include commercial fishing. The development of shellfish resources and fisheries management (e.g., growth of fish in rice paddies in Asia) are important aspects of this industry.

BOTANY

Botany is the study of plants. Plants were of paramount importance to early man; he depended upon them as sources of food, shelter, clothing, medicine, ornament, tools, and magic. Today it is known that, in addition to their practical and economic values, green plants are indispensable to all life on Earth: through the process of photosynthesis, plants transform energy from the sun into the chemical energy of food, which makes all life possible. A second unique and important capacity of green plants is the formation and release of oxygen as a by-product of photosynthesis. The oxygen of the atmosphere, so absolutely essential to many forms of life, represents the accumulation of over 3,500,000,000 years of photosynthesis by green plants.

Although the many steps in the process of photosynthesis have become fully understood only in recent years, even in prehistoric times man somehow recognized intuitively that some important relation existed between the sun and plants. Such recognition is suggested by the fact that, in primitive tribes and early civilizations, worship of the sun was often combined with the worship of plants.

Earliest man, like the other anthropoid mammals (e.g., apes, monkeys), depended totally upon the natural resources of his environment, which, until he developed methods for hunting, consisted almost completely of plants. The behaviour of pre-Stone Age man can be inferred by studying the botany of aboriginal peoples in various parts of the world. Isolated tribal groups in South America, Africa, and New Guinea, for example, have extensive knowledge about plants and distinguish hundreds of kinds according to their utility, as edible, poisonous, or otherwise important in their culture. They have developed surprisingly sophisticated systems of nomenclature and classification, which approximate the binomial system (i.e., generic and specific names) found in modern biology. The urge to recognize different kinds of plants and to give them names thus seems to be as old as the human race.

In time plants were not only collected by primitive man but also grown by him. This domestication resulted not only in the development of agriculture but also in a

greater stability of human populations that had previously been nomadic. From the settling down of agricultural peoples in places where they could depend upon adequate food supplies came the first villages and the earliest civilizations.

Because of the long preoccupation of man with plants, a large body of folklore, general information, and actual scientific data has accumulated, which has become the basis for the science of botany.

Historical background.

Theophrastus, a Greek philosopher who studied first with Plato and then became a disciple of Aristotle, is credited with founding botany. Only two of an estimated 200 botanical treatises written by him are known to science: originally written in Greek about 300 BC, they have survived in the form of Latin manuscripts, De causis plantarum and De historia plantarum. His basic concepts of morphology, classification, and the natural history of plants, accepted without question for many centuries, are now of interest primarily because of Theophrastus' independent and philosophical viewpoint.

Pedanius Dioscorides, a Greek botanist of the 1st century AD, was the most important botanical writer after Theophrastus. In his major work, an herbal in Greek, he described some 600 kinds of plants, with comments on their habit of growth and form as well as on their medicinal properties.

Unlike Theophrastus, who classified plants as trees, shrubs, and herbs, Dioscorides grouped his plants under three headings: as aromatic, culinary, and medicinal. His herbal, unique in that it was the first treatment of medicinal plants to be illustrated, remained for about 15 centuries the last word on medical botany in Europe.

From the 2nd century BC to the 1st century AD, a succession of Roman writers - Cato, Varro, Virgil, and Columella - prepared Latin manuscripts on farming, gardening, and fruit growing but showed little evidence of the spirit of scientific inquiry for its own sake that was so characteristic of Theophrastus. In the 1st century AD, Pliny the Elder, though no more original than his Roman predecessors, seemed more industrious as a compiler. His Historia naturalis - an encyclopaedia of 37 volumes, compiled from some 2,000 works representing 146

Roman and 327 Greek authors - has 16 volumes devoted to plants. Although uncritical and containing much misinformation, this work contains much information otherwise unavailable, since most of the volumes to which he referred have been destroyed.

The printing press revolutionized the availability of all types of literature, including that of plants. In the 15th and 16th centuries, many herbals were published with the purpose of describing plants useful in medicine. Written by physicians and medically oriented botanists, the earliest herbals were based largely on the work of Dioscorides and to a lesser extent on Theophrastus, but gradually they became the product of original observation. The increasing objectivity and originality of herbals through the decades is clearly reflected in the improved quality of the woodcuts prepared to illustrate these books.

In 1552 an illustrated manuscript on Mexican plants, written in Aztec, was translated into Latin by Badianus; other similar manuscripts known to have existed seem to have disappeared. Whereas herbals in China date back much further than those in Europe, they have become known only recently and so have contributed little to the progress of Western botany.

The invention of the optical lens during the 16th century and the development of the compound microscope about 1590 opened an era of rich discovery about plants; prior to that time, all observations by necessity had been made with the unaided eye. The botanists of the 17th century turned away from the earlier emphasis on medical botany and began to describe all plants, including the many new ones that were being introduced in large numbers from Asia, Africa, and America. Among the most prominent botanists of this era was Gaspard Bauhin, who for the first time developed, in a tentative way, many botanical concepts still held as valid. In 1665 Robert Hooke published, under the title Micrographia, the results of his microscopic observations on several plant tissues. He is remembered as the coiner of the word cell, referring to the cavities he observed in thin slices of cork; his observation that living cells contain sap and other materials too often has been forgotten. In the following decade, Nehemiah Grew and Marcello Malpighi

founded plant anatomy; in 1671 they communicated the results of microscopic studies simultaneously to the Royal Society of London, and both later published major treatises.

Experimental plant physiology began with the brilliant work of Stephen Hales, who published his observations on the movements of water in plants under the title Vegetable Staticks (1727). His conclusions on the mechanics of water transpiration in plants are still valid, as is his discovery—at the time a startling one - that air contributes something to the materials produced by plants. In 1774, Joseph Priestley showed that plants exposed to sunlight give off oxygen, and Jan Ingenhousz demonstrated, in 1779, that plants in the dark give off carbon dioxide. In 1804 Nicolas de Saussure demonstrated convincingly that plants in sunlight absorb water and carbon dioxide and increase in weight, as had been reported by Hales nearly a century earlier.

The widespread use of the microscope by plant morphologists provided a turning point in the 18th century - botany became largely a laboratory science. Until the invention of simple lenses and the compound microscope, the recognition and classification of plants were, for the most part, based on such large morphological aspects of the plant as size, shape, and external structure of leaves, roots, and stems. Such information was also supplemented by observations on more subjective qualities of plants, such as edibility and medicinal uses.

In 1753 Linnaeus published his master work, Species Plantarum, which contains careful descriptions of 6,000 species of plants from all of the parts of the world known at the time. In this work, which is still the basic reference work for modern plant taxonomy, Linnaeus established the practice of binomial nomenclature - that is, the denomination of each kind of plant by two words, the genus name and the specific name, as Rosa canina, the dog rose. Binomial nomenclature had been introduced much earlier by some of the herbalists, but it was not generally accepted; most botanists continued to use cumbersome formal descriptions, consisting of many words, to name a plant. Linnaeus for the first time put the contemporary knowledge of plants into an orderly system, with full

acknowledgment to past authors, and produced a nomenclatural methodology so useful that it has not been greatly improved upon. Linnaeus also introduced a "sexual system" of plants, by which the numbers of flower parts - especially stamens, which produce male sex cells, and styles, which are prolongations of plant ovaries that receive pollen grains—became useful tools for easy identification of plants. This simple system, though effective, had many imperfections. Other classification systems, in which as many characters as possible were considered in order to determine the degree of relationship, were developed by other botanists; indeed, some appeared before the time of Linnaeus. The application of the concepts of Charles Darwin (on evolution) and Gregor Mendel (on genetics) to plant taxonomy has provided insights into the process of evolution and the production of new species.

Systematic botany now uses information and techniques from all the subdisciplines of botany, incorporating them into one body of knowledge. Phytogeography (the biogeography of plants), plant ecology, population genetics, and various techniques applicable to cells - cytotaxonomy and cytogenetics - have contributed greatly to the current status of systematic botany and have to some degree become part of it. More recently, phytochemistry, computerized statistics, and fine-structure morphology have been added to the activities of systematic botany.

The 20th century has seen an enormous increase in the rate of growth of research in botany and the results derived therefrom. The combination of more botanists, better facilities, and new technologies, all with the benefit of experience from the past, has resulted in a series of new discoveries, new concepts, and new fields of botanical endeavour. Some important examples are mentioned below.

New and more precise information is being accumulated concerning the process of photosynthesis, especially with reference to energy-transfer mechanisms.

The discovery of the pigment phytochrome, which constitutes a previously unknown light-detecting system in plants, has greatly increased knowledge of the influence of both internal and external environment on the germination of seeds and the time of flowering.

Several types of plant hormones (internal regulatory substances) have been discovered - among them auxin, gibberellin, and kinetin - whose interactions provide a new concept of the way in which the plant functions as a unit.

The discovery that plants need certain trace elements usually found in the soil has made it possible to cultivate areas lacking some essential element by adding it to the deficient soil.

The development of genetical methods for the control of plant heredity has made possible the generation of improved and enormously productive crop plants.

The development of radioactive-carbon dating of plant materials as old as 50,000 years is useful to the paleobotanist, the ecologist, the archaeologist, and especially to the climatologist, who now has a better basis on which to predict climates of future centuries.

The discovery of alga-like and bacteria-like fossils in Precambrian rocks has pushed the estimated origin of plants on Earth to 3,500,000,000 years ago.

The isolation of antibiotic substances from fungi and bacteria-like organisms has provided control over many bacterial diseases and has contributed biochemical information of basic scientific importance as well.

Areas of study.

For convenience, but not on any mutually exclusive basis, several major areas or approaches are recognized commonly as disciplines of botany; these are morphology, physiology, ecology, and systematics.

Morphology.

Morphology deals with the structure and form of plants and includes such subdivisions as: cytology, the study of the cell; histology, the study of tissues; anatomy, the study of the organization of tissues into the organs of the plant; reproductive morphology, the study of life cycles; and experimental morphology, or morphogenesis, the study of development.

Physiology.

Physiology deals with the functions of plants. Its development as a subdiscipline has been closely interwoven with the development of other aspects of botany,

especially morphology. In fact, structure and function are sometimes so closely related that it is impossible to consider one independently of the other. The study of function is indispensable for the interpretation of the incredibly diverse nature of plant structures. In other words, around the functions of the plant, structure and form have evolved. Physiology also blends imperceptibly into the fields of biochemistry and biophysics, as the research methods of these fields are used to solve problems in plant physiology.

Ecology.

Ecology deals with the mutual relationships and interactions between organisms and their physical environment. The physical factors of the atmosphere, the climate, and the soil affect the physiological functions of the plant in all its manifestations, so that, to a large degree, plant ecology is a phase of plant physiology under natural and uncontrolled conditions; in fact, it has been called "outdoor physiology." Plants are intensely sensitive to the forces of the environment, and both their association into communities and their geographical distribution are determined largely by the character of climate and soil. Moreover, the pressures of the environment and of organisms upon each other are potent forces, which lead to new species and the continuing evolution of larger groups.

Systematics.

Systematics deals with the identification and ranking of all plants; it includes classification and nomenclature (naming) and enables the botanist to comprehend the broad range of plant diversity and evolution.

Other subdisciplines.

In addition to the major subdisciplines, several specialized branches of botany have developed as a matter of custom or convenience. Among them are bacteriology, the study of bacteria; mycology, the study of fungi; algology or phycology, the study of algae; bryology, the study of mosses and liverworts; pteridology, the study of ferns and their relatives; and paleobotany, the study of fossil plants. Palynology is the study of modern and fossil pollen and spores, with particular reference to their identification; plant pathology deals with the diseases of plants; economic

botany deals with plants of practical use to man; and ethnobotany covers the use of plants by aboriginal peoples, now and in the distant past.

Botany also relates to other scientific disciplines in many ways, especially to zoology, medicine, microbiology, agriculture, chemistry, forestry, and horticulture, and specialized areas of botanical information may relate closely to such humanistic fields as art, literature, history, religion, archaeology, sociology, and psychology.

Fundamentally, botany remains a pure science, including any research into the life of plants and limited only by man's technical means of satisfying his curiosity. It has often been considered an important part of a liberal education, not only because it is necessary for an understanding of agriculture, horticulture, forestry, pharmacology, and other applied arts and sciences, but also because an understanding of plant life is related to life in general.

Because man has always been dependent upon plants and surrounded by them, he has woven them into his designs, into the ornamentation of his life, even into his religious symbolism. A Persian carpet and a bedspread from a New England loom both employ conventional designs derived from the forms of flowers. Medieval painters and great masters of the Renaissance represented various revered figures surrounded by roses, lilies, violets, and other flowers, which symbolized chastity, martyrdom, humility, and other Christian attributes.

Methods in botany.

Morphological aspects.

The invention of the compound microscope provided a valuable and durable instrument for the investigation of the inner structure of plants. Early plant morphologists, especially those studying cell structure, were handicapped as much by the lack of adequate knowledge of how to prepare specimens as they were by the imperfect microscopes of the time. A revolution in the effectiveness of microscopy occurred in the second half of the 19th century with the introduction of techniques for fixing cells and for staining their component parts. Before the development of these techniques, the cell, viewed with the microscope, appeared

as a minute container with a dense portion called the nucleus. The discovery that parts of the cell respond to certain stains made observation easier. The development of techniques for preparing tissues of plants for microscopic examination was continued in the 1870s and 1880s and resulted in the gradual refinement of the field of nuclear cytology, or karyology. Chromosomes were recognized as constant structures in the life cycle of cells, and the nature and meaning of meiosis, a type of cell division in which the daughter cells have half the number of chromosomes of the parent, was discovered; without this discovery, the significance of Mendel's laws of heredity might have gone unrecognized. Vital stains, dyes that can be used on living material, were first used in 1886 and have been greatly refined since then.

Improvement of the methodology of morphology has not been particularly rapid, even though satisfactory techniques for histology, anatomy, and cytology have been developed. The embedding of material in paraffin wax, the development of the rotary microtome for slicing very thin sections of tissue for microscope viewing, and the development of stain techniques are refinements of previously known methods. The invention of the phase microscope made possible the study of unfixed and unstained living material - hopefully nearer its natural state. The development of the electron microscope, however, has provided the plant morphologist with a new dimension of magnification of the structure of plant cells and tissues. The fine structure of the cell and of its components, such as mitochondria and the Golgi apparatus, have come under intensive study. Knowledge of the fine structure of plant cells has enabled investigators to determine the sites of important biochemical activities, especially those involved in the transfer of energy during photosynthesis and respiration. The scanning electron microscope, a relatively recent development, provides a three-dimensional image of surface structures at very great magnifications.

For experimental research on the morphogenesis of plants, isolated organs in their embryonic stage, clumps of cells, or even individual cells are grown. One of the most interesting techniques developed thus far permits the growing of plant tissue

of higher plants as single cells; aeration and continuous agitation keep the cells suspended in the liquid culture medium.

Physiological aspects.

Plant physiology and plant biochemistry are the most technical areas of botany; most major advances in physiology also reflect the development of either a new technique or the dramatic refinement of an earlier one to give a new degree of precision. Fortunately, the methodology of measurement has been vastly improved in recent decades, largely through the development of various electronic devices. The phytotron at the California Institute of Technology represents the first serious attempt to control the environment of living plants on a relatively large scale; much important information has been gained concerning the effects on plants of day length and night length and the effects on growth, flowering, and fruiting of varying night temperatures. Critical measurements of other plant functions have also been obtained.

Certain complex biochemical processes, such as photosynthesis and respiration, have been studied stepwise by immobilizing the process through the use of extreme cold or biochemical inhibitors and by analyzing the enzymatic activity of specific cell contents after spinning cells at very high speeds in a centrifuge. The pathways of energy transfer from molecule to molecule during photosynthesis and respiration have been determined by biophysical methods, especially those utilizing radioactive isotopes.

An investigation of the natural metabolic products of plants requires, in general, certain standard biochemical techniques - e.g., gas and paper chromatography, electrophoresis, and various kinds of spectroscopy, including infrared, ultraviolet, and nuclear magnetic resonance. Useful information on the structure of the extremely large cellulose molecule has been provided by X-ray crystallography.

Ecological aspects.

When plant ecology first emerged as a subscience of botany, it was largely descriptive; today, however, it has become a common meeting ground for all the plant sciences, as well as for other sciences. In addition, it has become much more

quantitative. As a result, the tools and methods of plant ecologists are those available for measuring the intensity of the environmental factors that impinge on the plant and the reaction of the plant to these factors. The extent of the variability of many physical factors must be measured. The integration and reporting of such measurements, which cannot be regarded as constant, may therefore conceal some of the most dynamic and significant aspects of the environment and the responses of the plant to them. Because the physical environment is a complex of biological and physical components, it is measured by biophysical tools. The development of electronic measuring and recording devices has been crucial for a better understanding of the dynamics of the environment. Such devices, however, produce so much information that computer techniques must be used to reduce the data to meaningful results.

The ecologist, concerned primarily with measuring the effect of the external environment on a plant, adapts the methodology of the plant physiologist to field conditions.

The plant sociologist, on the other hand, is concerned with both the relation of different kinds of plants to each other and the nature and constitution of their association into natural communities. One widely used technique in this respect is to count the various kinds of plants within a standard area in order to determine such factors as the percentage of ground cover, dominance of species, aggressiveness, and other characteristics of the community.

In general, the plant sociologist has relatively few quantitative factors to measure and must therefore take a subjective and intuitive approach, which, nevertheless, gives extremely useful results and some degree of predictability.

Some ecologists are most concerned with the inner environment of the plant and the way in which it reacts to the external environment. This approach, which is essentially physiological and biochemical, is useful for determining energy flow in ecosystems. The physiological ecologist is also concerned with evaluating the adaptations that certain plants have made toward survival in a hostile environment.

In summary, the techniques and methodology of plant ecology are as diverse and as varied as the large number of sciences that are drawn upon by ecologists. Completely new techniques, although few, are important; among them are techniques for measuring the amount of radioactive carbon-14 in plant deposits up to 50,000 years old. The most important new method in plant ecology is the rapidly growing use of computer techniques for handling vast amounts of data. Furthermore, modern digital computers can be used to simulate simple ecosystems and to analyze real ones.

Taxonomic aspects.

Experimental research under controlled conditions, made possible by botanical gardens and their ranges of greenhouses and controlled environmental chambers, has become an integral part of the methodology of modern plant taxonomy.

A second major tool of the taxonomist is the herbarium, a reference collection consisting of carefully selected and dried plants attached to paper sheets of a standard size and filed in a systematic way so that they may be easily retrieved for examination. Each specimen is a reference point representing the features of one plant of a certain species; it lasts indefinitely if properly cared for, and, if the species becomes extinct in nature—as hundreds have - it remains the only record of the plant's former existence. The library is also an essential reference resource for descriptions and illustrations of plants that may not be represented in a particular herbarium.

One of the earliest methods of the taxonomist, the study of living plants in the field, has benefitted greatly by fast and easy methods of transportation; botanists may carry on fieldwork in any part of the world and make detailed studies of the exact environmental conditions under which each species grows.

During the present century, many new approaches have been applied to the elucidation of problems in systematic botany. The transmission electron microscope and the scanning electron microscope have added to the knowledge of plant morphology, upon which classical taxonomy so much depends.

Refined methods for cytological and genetical studies of plants have given the taxonomist new insights into the origin of the great diversity among plants, especially the mechanisms by which new species arise and by which they then maintain their individuality in nature. From such studies have arisen further methods and also the subdisciplines of cytotaxonomy, cytogenetics, and population genetics.

Phytochemistry, or the chemistry of plants, one of the early subdivisions of organic chemistry, has been of great importance in the identification of plant substances of medicinal importance. With the development of new phytochemical methods, new information has become available for use in conjunction with plant taxonomy; thus has arisen the modern field of chemotaxonomy, or biochemical systematics. Each species tends to differ to some degree from every other species, even in the same genus, in the biochemistry of its natural metabolic products. Sometimes the difference is subtle and difficult to determine; sometimes it is obvious and easily perceptible. With new analytical techniques, a large number of individual compounds from one plant can be identified quickly and with certainty. Such information is extremely useful in adding confirmatory or supplemental evidence of an objective and quantitative nature. An interesting by-product of chemical plant taxonomy has resulted in understanding better the restriction of certain insects to specific plants.

Computer techniques have recently been applied to plant taxonomy to develop a new field, numerical taxonomy, or taximetrics, by which relationships between plant species or those within groups of species are determined quantitatively and depicted graphically. Another method measures the degree of molecular similarity of deoxyribonucleic acid (DNA) molecules in different plants. By this procedure it should be possible to determine the natural taxonomic relationships among different plants and plant groups by determining the extent of the relationship of their DNA's: closely related plants will have more similarities in their DNA's than will unrelated ones.

MICROBIOLOGY

The 17th-century discovery that living forms exist that are invisible to the naked eye was a dramatic one in man's history, for, from the 13th century onward, it had been postulated that "invisible" organisms were responsible for decay and disease. The word microbe was coined in the latter quarter of the 19th century to describe these organisms, all of which were thought to be related. As microbiology eventually developed into a separate science, it was found that microbes comprise a very large group of extremely diverse organisms; thus, microbiology became subdivided into various disciplines - e.g., bacteriology, protozoology, and virology. The diversity of microbes, or microorganisms as they are now commonly called, has meant that it is almost impossible for one person to be knowledgeable in all of the disciplines grouped under microbiology.

Microbiology involves the identification of microorganisms and the study of their structure and function. It encompasses the study of bacteria, rickettsiae, small fungi (e.g., yeasts and molds), algae, and protozoans, as well as problematical forms of life such as viruses. Because of the difficulty of assigning plant or animal status to microorganisms - some are plantlike, others animal-like - they are sometimes considered a separate group called protists. Microbes can also be divided into procaryotes, which have a primitive and dispersed kind of nuclear material - i.e., the blue-green algae, bacteria, and rickettsiae - and eucaryotes, which display a distinct nucleus bounded by a membrane. Such are small algae other than the blue-greens, yeasts and molds, and protozoans. (All higher organisms are eucaryotes.)

Man's daily life is interwoven inextricably with microorganisms. They abound in the soil, in the seas, and in the air. Everywhere abundant, although usually unnoticed, microorganisms provide ample evidence of their presence, sometimes unfavourably, as when they cause decay of objects valued by man or generate disease, and sometimes favourably, as when they ferment alcohol to wine and beer, raise bread, flavour cheeses, and create other dairy products from milk. Microorganisms are of incalculable value in nature, causing the disintegration of

animal and plant remains and converting them to gases and minerals that can be recycled in other organisms.

Historical background.

Microbiology can be said to have begun with the development of the microscope. Although others may have seen microbes before him, Antonie van Leeuwenhoek, a Dutch draper whose hobby was lens grinding and microscope making, was the first to provide proper descriptions of his observations, which included protozoans from the guts of animals and bacteria from teeth scrapings. His descriptions and drawings were excellent because his lenses were of an exceptional quality. Leeuwenhoek conveyed his findings in a series of letters to the British Royal Society during the mid-1670s.

Although his observations stimulated much interest, no one made a serious attempt either to repeat or to extend them. Leeuwenhoek's "animalcules," as he called them, thus remained mere oddities of nature to the scientists of his day, and enthusiasm for the study of microbes gained ground slowly. It was only later, during the 18th-century revival of a long-standing controversy about whether or not life can develop out of nonliving material, that the significance of microorganisms in the scheme of nature and in the health and welfare of man became evident.

The early Greeks believed that living things could originate from nonliving matter; the goddess Gea was credited with creating life from stones. Although Aristotle discarded this notion, he still held that animals could arise spontaneously from other unlike organisms or from soil. His influence regarding this concept of spontaneous generation was still felt as late as the 17th century. Toward the end of the 17th century, a chain of observations, experiments, and arguments began that dealt a deathblow to the idea that life could be generated from nonlife. It was an involved series of events, with the forces of personality and strong wills often obscuring the facts.

Although Francesco Redi, an Italian naturalist, disproved that higher forms of life could originate spontaneously, proponents of the concept claimed that microbes were different and did indeed arise in this way. Such illustrious names as John Needham, Lazzaro Spallanzani, Franz Schultze, and Theodor Schwann figured in the debates.

It remained for Louis Pasteur to settle the matter. He proved in a series of masterful experiments that only preexisting microbes could give rise to other microbes - at least under current earthly conditions (that life arose spontaneously from nonlife at some earlier time, under appropriate physical and chemical conditions, is an undisputed postulate of chemical evolution).

Regarding microbes and disease, Girolamo Fracastoro, an Italian scholar, advanced the notion as early as the mid-1500s that contagion is an infection that passes from one thing to another. The "thing" that is passed along eluded discovery until the late 1800s, when the work of many scientists, Pasteur foremost among them, determined the role of bacteria in fermentation and disease. Robert Koch, a German physician, defined the procedure for proving that a specific organism causes a specific disease.

The foundation of microbiology was securely laid during the period from about 1880 to 1900. The students of Pasteur, Koch, and others discovered in rapid succession a host of bacteria capable of causing specific diseases (pathogens) and elaborated an extensive armamentarium of techniques and laboratory procedures for revealing the ubiquity, diversity, and power of microbes.

All of these developments occurred in Europe. Not until the early 1900s did microbiology become established in America. Many of the microbiologists who worked in America at this time either had studied under Koch or at the Pasteur Institute, in Paris. All microbiologists of the early 20th century, however, were influenced by such men as Koch. Once established in America, microbiology flourished, especially with regard to such related disciplines as biochemistry and genetics.

Since the 1940s, microbiology has experienced an extremely productive period, during which many disease-causing microbes have been identified and methods to control them have been developed. Microorganisms have also been effectively utilized in industry; their activities have been channelled so that valuable products of commerce and agricultural benefits result.

The study of microorganisms has also advanced man's knowledge of living things. Microbes provide easy-to-work-with material for studying the complex processes of life; e.g., metabolism. Correlated with the intensive probing into the functions of microbes have been numerous, and often unexpected, dividends that can be applied to solving existing problems. Knowledge of the basic metabolism of a pathogenic bacterium, for example, often leads to a means for controlling the pathogen. Nutritional requirements of bacteria thus may be of value in combatting an infection.

Areas of study.

The study of bacteria, which were among the first objects of microbiological study, is called bacteriology. Various subdisciplines deal exclusively with particular microorganisms.

From another standpoint microbiology can be subdivided into theoretical, or pure, microbiology, and practical, or applied, microbiology. The latter can be further subdivided according to specialties, such as medical, industrial, agricultural, food, and dairy microbiology.

Interdisciplinary work.

The science of microbiology has been influenced by, and in turn has influenced, other sciences. Microorganisms are no longer the exclusive concern of microbiologists. Biochemists, geneticists, cytologists, and molecular biologists have discovered the value of microbes as experimental tools in the study of such fundamental biological processes as metabolism, photosynthesis, enzyme action, gene action, and population dynamics. Microorganisms are well suited to such

uses; they represent a vast range of metabolic types, and genetic changes are correlated with their rapid proliferation. In addition, they can subsist on relatively simple inorganic nutrients and, because they multiply rapidly, are available in extremely large numbers in a relatively short period of time. They are also relatively easy to maintain and handle under laboratory conditions.

Bacteriology.

Modern and accurate knowledge of the forms of bacteria dates from the researches of the German botanist F.J. Cohn, the chief results of which were published at various periods between

1853 and 1892. Cohn's classification of bacteria, published in 1872 and extended in 1875, dominated the study of these organisms thereafter. While various observers added to the knowledge of the structure of bacteria, others laid the foundation of what is known about the relations of bacteria to fermentation and disease. When Pasteur showed in 1857 that lactic acid fermentation depends upon the presence of an organism, it was already known that fermentation and putrefaction are intimately connected with the presence of organisms in the air. In 1862 Pasteur placed beyond reasonable doubt the fact that production of ammonia by the fermentation of urea is caused by the action of a minute bacterium, named in 1874 Micrococcus ureae.

After the introduction of bacteriological techniques (see below Cultivation of microorganisms) came the isolation of many bacteria. It was discovered in 1882, for example, that a bacillus (a rod-shaped bacterium) is the cause of glanders, a disease of horses. In 1883 Koch isolated the organism of Asiatic cholera, and the same year that of diphtheria was found. In 1885 the tetanus bacillus was observed in pus produced in mice and rabbits inoculated with soil; only in 1889, however, did the Japanese bacteriologist S. Kitasato discover the way in which the organisms could be cultivated (they grow only in the absence of oxygen). W.D. Miller, a U.S. dentist, studied the microorganisms of the human mouth in the 1880s, noting their possible relationship to the decay of teeth.

Pasteur and his associates found that animals vaccinated with a specially cultivated anthrax bacillus showed immunity to disease when reinoculated with the deadly wild form. These findings were destined to lead to a study of the principles of immunity, which underlie the prevention and treatment of disease by vaccines and immune serum. Questions relating to causes and to the nature of changes occurring both in the bacteria and in the host, as well as the development of immunity in the latter, continue to be subjects of great interest and importance.

While investigations on infectious diseases and immunity were under way, it became apparent that other activities of bacteria also are of importance to man. In 1878, only two years after Koch announced the discovery of the anthrax bacillus, the U.S. botanist T.J. Burrill discovered the bacterial cause of fire blight in pears, thereby establishing that certain plant diseases are caused by bacteria. The importance of some bacteria in soil and their contribution to soil fertility was recognized in the 1880s and 1890s. The significance of bacterial activities in many aspects of the dairy industry was recognized at about the same time.

The further application of bacterial activities to industrial processes, other than early studies on alcoholic and lactic acid fermentations, began later. Thus, within the span of a few decades the study of bacteria had progressed enormously, and by the early 20th century it was recognized that the activities of bacteria bear an intimate relation to many aspects of human activity.

The original interest of bacteriologists centred upon what bacteria do. This interest has broadened to include the study of the bacteria themselves: what they are, their relationships to each other, and their relationships to other organisms.

Protozoology and others.

The name Protozoa is derived from the Greek words meaning "first animal," and, indeed, protozoans are sometimes considered the most simple of all animals. First observed by Leeuwenhoek in the 17th century, protozoans are almost as ubiquitous as bacteria. The study of protozoans at one time centred on the parasites that cause malaria and sleeping sickness, thereby stimulating research in tropical medicine.

Improved laboratory techniques for cultivating this diverse group of organisms have made them valuable tools in many types of investigations - from physiological studies, such as cytoplasmic motion in Amoeba, to ecological studies involving their role in the food chain of the oceans.

The term fungi, Latin for "mushroom," applies not only to mushrooms but to all of the large and diverse group of plantlike organisms to which the mushrooms belong. The fungi of most interest in microbiology are the yeasts and molds. Many molds are studied because of their economic importance. Mucor, for example, not only causes spoilage of foods such as vegetables and fruits but also is used to manufacture some cheeses.

Some molds form unique relationships with other organisms; lichens, for instance, are composed of both algae and fungi. The unusual nature of lichens was not discovered until after the invention of the microscope. The name lichen appeared long before, however, in the writings of Theophrastus, a disciple of Aristotle, about the 4th century BC. At that time, and for many centuries thereafter, lichens were confused with mosses and the hepatics. Early students thought that the green cells, now known to be an alga, were specialized fungal structures. Not until 1867 were the green cells in lichens indentified as algae; at the same time it was announced that the fungal cells were parasitic upon them. The introduction of pure culture techniques permitted the lichen constituents to be isolated and grown apart from each other. In 1873 the German botanist H.A. De Bary suggested the name symbiosis to describe the mutual benefit between the components of lichens.

Once objects of interest for their curious form, slime molds - as their individual nutritional, environmental, and genetic characteristics have become better understood - have become objects useful in teaching and research. The creeping sheet (plasmodium), or vegetative stage of a slime mold, has the characteristics of a primitive animal and resembles a primitive amoeba. The reproductive stage (sporangium) has the characteristics of a mold. The double life of these organisms is reflected in the name Mycetozoa, meaning fungus animals, which was coined by De Bary in 1858.

Algae, a heterogeneous group of primitive organisms, are of great interest to all biologists because they are capable of photosynthesis and are of evolutionary interest as well. Intensive research is currently directed toward the small forms making up much of the free-floating microscopic life in water (phytoplankton) because they provide food for aquatic animals and thus are of great importance. Many freshwater algae produce undesirable tastes and odours, and studies have been aimed at eliminating them from domestic water supplies. Although algae occasionally form mats on water surfaces, sometimes causing suffocation of aquatic life, heavy growth of certain species also has been found to reduce water hardness and to remove the salts found in brackish water, thereby making the water more suitable for human consumption. Algae are also being studied for their potential value as a food source for man.

Virology and others.

Viruses comprise a heterogeneous assemblage of self-reproducing agents, smaller than the microscopically visible bacteria, that multiply only within living susceptible cells. Certain diseases caused by viruses have been known since the late 1700s, when the British physician Edward Jenner developed a vaccine from material isolated from cowpox lesions. He did not see the causative agent. Pasteur, in the mid-1850s, developed an attenuated strain (i.e., one made less virulent) of the virus that causes rabies, but he was not then aware of the viral nature of the disease. An associate of Pasteur, Charles Chamberland, discovered that bacteria would not pass through a porcelain filter but that the causative agent of rabies did. The term filter-passing, or filterable, virus was used to describe these agents.

The Russian bacteriologist D. Ivanovski in 1892 found that a filtrate of sap from tobacco plants infectedwith mosaic disease could be used to transfer infection to healthy plants; the Dutch microbiologist M.W. Beirjerinck later confirmed the work. The further finding in 1902 that the agent of foot-and-mouth disease of cattle is also filterable made clear the fact that the agents cause disease in animals as well

as plants. By 1930 the term filterable had been dropped, and virus is routinely used by microbiologists as a name for these agents.

Nobel Prize winner Max Theiler found in 1951 a method for attenuating virulent yellow fever virus; the technique has since been modified to produce vaccines against other viral diseases. Thus, viruses have been studied intensively in recent years not only because they provide important information about life processes but also because vaccines against them are constantly sought.

The study of rickettsiae, which resemble very small bacteria but grow only in susceptible cells, is intimately associated with those forms causing human disease. Elucidation of the microbes causing such diseases as epidemic typhus, Rocky Mountain spotted fever, and scrub typhus began in 1906 with the work of Howard Taylor Ricketts, for whom the organisms are named; he described the organism of spotted fever. In 1910 Ricketts and an associate described the organism of typhus fever, which in 1916, was named Rickettsia prowazekii in honour of Ricketts and Stanislas von Prowazek, both of whom died of the disease. Research continues as better methods for prevention, control, and treatment of rickettsial diseases are sought.

General trends.

One of the more current studies involving microorganisms is their possible occurrence in outer space and on planets other than Earth. A branch of exobiology, space microbiology, includes the investigation of microbes as providers of food and oxygen in the closed environment of spaceships.

A less positive development of microbiology has been biological warfare - the selection and cultivation of microbes as weapons of war, to cause disease or injury to domestic plants, animals, and man.

Although most of the subdisciplines of microbiology are directed at the occurrence of microbes in specific environments, the study of gnotobiotics is concerned with the exclusion of microbes except those involved in any given experiment. Such

germfree organisms are important research tools in investigating parasitic diseases, immunological processes, nutrition, stress, shock, and aging.

Methods in microbiology.

The study of microbiology is channelled in two major directions. Pure microbiology - that is, the study of a particular group of microorganisms in order to learn about their morphology, physiology, taxonomy, occurrence, variation, heredity, and evolution - seeks to understand the nature of microbes. Applied microbiology, on the other hand, is motivated by the desire to exploit the effects of microorganisms that are of benefit to man and to control the activities of those that are harmful.

The methods used to study bacteria have been widely adapted to the study of other microorganisms; indeed, the techniques employed in the microbiological sciences probably have more in common than do the origin and evolution of the organisms themselves. For this reason, the term microbiology is often considered to be synonymous with bacteriology. Similarly, the term microorganism no longer refers only to microscopic organisms. With the adoption of bacteriological methods for the study of many types of organisms, the meaning of the term microorganism has gradually been extended to include any organism that can grow and be studied using the methods originally developed for bacteria. The study of microbiology, therefore, encompasses many organisms that are not microscopic (e.g., molds) but usually does not include some that are (e.g., rotifers).

That invisibly small organisms exist was believed from early times. Lucretius, who expressed an atomic view of matter, wrote about AD 75 that even the plague must be caused by a kind of atom.

He considered the atoms, or seeds, as lifeless, however. The writings of Lucretius influenced 16th-century scientists, but the means for actually observing microorganisms - a magnifying lens - was not developed until the 17th century by Galileo.

The convex lens of Leeuwenhoek allowed a greatly enlarged image to be seen, enabling him to see protozoans, filamentous fungi, and yeasts. In 1676, using his simple lens, which was capable of magnifying 280 times, he first saw bacteria. A few years before Leeuwenhoek's observation of bacteria, Robert Hooke, the father of the cell theory, had observed filamentous fungi (1667) through a compound microscope - one with two lenses, which gives a larger image than that given by a simple lens.

By 1786 various microorganisms, including bacteria, had been viewed through the compound microscope. The compound microscope of today is widely used; it differs from those of the 18th century mainly in the addition of a device, called a condenser, for lighting the object being viewed with a wide cone of light. A condenser is necessary to provide sufficient light when large magnifications are used.

Of the types of microscopy now used in microbiology, most utilize light microscopes; magnification is obtained by a system of optical lenses. The types of light microscopes include the commonly used bright field, in which the background is brightly lighted, the objects studied are dark, and the power of magnification is about 1,000; the dark field, in which the background is dark and the objects studied are bright, a phenomenon particularly useful for examining unstained organisms suspended in fluid; the ultraviolet, the greatest advantage of which is that magnifications two to three times greater than those obtained with the light microscope can be achieved because ultraviolet light has a shorter wavelength than visible light; and phase contrast, in which controlled illumination is obtained by the use of special equipment that enables the refraction, or bending, of light passing from one material to another to be seen, thereby revealing differences in cells not discernible with other microscopic methods. Fluorescence microscopy utilizes the ultraviolet microscope; a chemical substance with the property to absorb ultraviolet waves and emit visible ones is used with a mixture of microorganisms, some of which take up the substance, and thus can be distinguished from those that do not. Still another type of microscopy, electron

microscopy, allows very great magnification; waves of electrons (negatively charged particles), rather than light, and magnetic fields, rather than lenses, are used to achieve an image. Although electron microscopy has the advantage of great magnification, it also has several limitations, most important of which is that the material must be dry. This not only eliminates the possibility of studying living specimens but also raises the possibility that the drying process may alter the characteristics of the specimen.

Cultivation of microorganisms.

Microorganisms growing on a nutrient medium are referred to as a culture. Early experiments (1776) by Spallanzani, begun as the result of the previously mentioned controversy over whether or not life could develop out of lifeless matter, laid the foundation for the technique of sterile culture. This involves first freeing a suitable medium of microorganisms by heating and, second, keeping the medium sterile - i.e., keeping microorganisms out of the medium. In his experiments, Spallanzani boiled various kinds of seeds in a flask and stoppered it. After a few days, many organisms could be found in the flask; Spallanzani distinguished the larger ones, which were destroyed by boiling for one-half minute, and microbes, which survived boiling and developed even after the flask had been sealed. Eventually, he discovered that, after boiling sealed flasks for as long as 45 minutes, no microorganisms developed.

The objection then voiced to Spallanzani's work - that the quality of the air in the flasks had, by heating, been rendered unable to support life - was overcome later (1836) when air was passed into the flask after having been slowly drawn through solutions containing sulfuric acid. The meat extract in the flask remained uncontaminated because the microorganisms in the air were killed by its passage through the solutions. In 1853 it was discovered that flasks did not have to be sealed after boiling but could be closed with a plug of cotton, which effectively filtered the incoming air; this procedure is still used. Because many microorganisms can produce bodies (spores) that are extremely heat-resistant,

mediums for the growth of microorganisms are usually sterilized by heating under pressure to 120 C (250 F) for 15 or 20 minutes.

The Scottish surgeon Joseph Lister contributed to culture methods in microbiology with the introduction in 1878 of the dilution method. During studies probably concerning the souring of milk, he used sterile water to dilute a small amount of milk, then diluted some of this mixture with more sterile water. He continued the dilution process until a sample of the milk and sterile water mixture no longer caused the milk to sour. The last dilution that resulted in souring thus contained a minimum number of organisms, and Lister considered it a pure culture - that is, containing only one species of organism. The dilution method has become an essential part of pure-culture methods.

The dilution method was modified in 1896 by a Danish microbiologist, Emil Hansen, who studied yeasts. He added one drop of a yeast culture to the first of a series of small drops of sterile water, then removed a drop of the culture and added it to the second drop of water. Eventually, he obtained a drop containing just one yeast cell.

Koch also made a significant contribution to culture techniques with his development of a simple method for obtaining pure cultures of bacteria. He added a solidifying agent (gelatin) to a nutrient medium. After the medium had been heat sterilized and partially cooled, he added some microorganisms. By cooling the solution further and spreading the organisms before the gelatin solidified, he was able to isolate the microorganisms from each other, with the result that each organism gave rise to a separate colony, or crop of cells.

In 1883 a woman in Koch's laboratory, Frau Hesse, further improved the technique by substituting agar-agar for the gelatin. Unlike gelatin, agar-agar can be liquefied by only a few microorganisms and does not provide a food source, thus allowing better control of the nutrient content of the medium. Silica gel also has been used as a solidifying agent.

One final advance in culture technique was developed in the 1890s by M.W. Beirjerinck and by the Russian microbiologist Sergei Winogradsky. They selected

a medium that favours the growth of one organism over another. If an organism, one from the soil, for example, is capable of utilizing a certain substance (e.g., a specific sugar), growth of the organism can be induced by using the substance in the medium. Only the organisms capable of attacking the substance will grow in great numbers. Successive transfers of a relatively small number of organisms (inoculum) to a fresh medium eventually will result in a culture strongly enriched with the organism that utilizes the desired substance.

After a species of microorganism has been obtained as a pure culture, it is necessary to maintain it alive and as a pure culture. Microbiology laboratories in schools, universities, and industry usually maintain collections of pure cultures of the particular species they use. Various organizations throughout the world maintain pure cultures of microorganisms; such collections are important in that they make available pure cultures of species when needed.

The Kral Collection in Prague, established in 1900, was the first known culture collection. The names and locations of a few of the other collections are: the American Type Culture Collection (Rockville, Maryland), the Japanese Type Culture Collection (Tokyo), and the (British) National Collection of Type Cultures (London). Other collections also exist worldwide to serve particular needs. A section on culture collections was established in 1966 by the International Association of Microbiological Societies in order to promote the exchange of information among culture collections.

Whenever a microbiologist proposes a new species, he provides one or more of the national culture collections with a pure culture of the species.

Staining techniques.

Until the latter third of the 19th century, microorganisms were observed only in the natural state; the similarity of their refractive index to water and their small size, however, made them difficult to see when viewed through a microscope. Thus, the discovery of a method that would allow them to be seen more easily was an important development.

By 1875 the German pathologist Carl Weigert, using techniques developed more than a decade earlier for staining animal tissues with dyes, had found that dead ("fixed") bacteria would become heavily stained with the dye picrocarmine. Other dyes were then introduced - e.g., methylene blue, fuchsin, and crystal violet. Hans Christian Gram, a Danish bacteriologist, discovered in 1884 a simple procedure for placing bacteria into either a gram-negative class or a gram-positive class, depending on whether or not the organisms retained crystal violet when treated in a specific way. Although the gram stain is used mostly with bacteria, other microorganisms also show a reaction. Yeasts and actinomycetes, for example, are gram-positive; rickettsiae are gram-negative. Stained preparations now are used particularly to observe structural features of microorganisms.

Twentieth-century developments in microbiological methods have depended largely upon the techniques of other disciplines - namely, biochemistry, physiology, organic chemistry, and physics.

The gas exchange of respiring microorganisms, called the respiratory quotient (the ratio of carbon dioxide produced to oxygen consumed), is measured by following the pressure change of respiring organisms in a closed vessel. Other processes - e.g., fermentation - can also be studied using this method.

The transfer of hydrogen ions (positively charged atoms) in microorganisms has been studied by adding to the organisms or extracts prepared from them a substance (e.g., the dye methylene blue) the colour of which changes when hydrogenated. Many other dyes behave in this way.

The techniques used to break up microorganisms include such biochemical techniques as alternate freezing and thawing, prolonged grinding, vigorous shaking with glass beads, and exposure to supersonic vibration; certain microorganisms (e.g., yeast) are especially difficult to disrupt because of their resistant cell walls.

The apparatus of modern physics has greatly aided the advance of virology. About 1930 the steady development of new technical methods for studying the physical, chemical, and biological properties of microorganisms completely changed the outlook of biologists toward viruses. An important advance in this new approach

was the development of graded filter membranes of known pore size. This created for the first time the opportunity to assess the actual size of virus particles.

Another important advance of the 1930s was the growth of the virus of fowl pox in the tissues of a developing chick embryo. Methods for the growth in the laboratory of rickettsiae are similar in that living host cells are necessary. In 1949, tissue culture methods, long used in half-hearted fashion for the cultivation of viruses, were shown to be suitable for the growth of poliomyelitis viruses. This not only opened the way to immunization against polio but supplied a method by which many new types of virus could be isolated.

PHYSIOLOGY AND ITS IMPORTANCE

Physiology (Gk physis nature, logos science) is the science that deals with the study of the vital activity of an integral living organism and its separate parts: cells, tissues, organs, and functional systems. The aim of physiology is to reveal the mechanisms by which functions of the living organism are performed, to establish their interconnection, regulation, adaptation to the external environment, their origin and development.

The physiological regularities are based on the data on the macro- and microscopic structure of organs and tissues and on the biochemical and biophysical processes taking place in cells, organs, and tissues.

Physiology is concerned with the synthesis of the factual information provided by anatomy, histology, cytology, molecular biology, biochemistry, biophysics, and other sciences and unites it into a single system of knowledge. Thus, physiology is the science which rests upon a systemic approach that implies study of the organism and all its elements as systems. The systemic approach enables an investigator in the first place to reveal the integrity of the object and the mechanisms that provide for this integrity, to discover the diverse types of connections in a complex object and to bring them together into a single theoretical pattern.

PHYSIOLOGY AND MEDICINE

By uncovering the principal mechanisms that ensure the existence of an integral organism and its interaction with the external environment, physiology enables one to elucidate and investigate the causes, conditions, and character of these mechanisms' impairment during illness. Knowledge of physiology helps determine the ways and methods for exerting influence on the organism to normalize its functions and therefore to restore health. For this reason, physiology may be considered a theoretical foundation of medicine, and, therefore, physiology and medicine are inseparable. The physician can evaluate the severity of a disease by the degree of functional disturbances, i.e. by the value to which certain

physiological functions have deviated from the norm. At present, all abnormalities in the organism's functioning may be measured and evaluated quantitatively. The functional (physiological) studies form the basis of the clinical diagnosis and are a method for evaluating the effectiveness of treatment and prognosis of diseases. The task set before the physician who examines the patient and establishes the degree to which physiological functions have been impaired is to bring these functions to the norm.

The importance of physiology for medicine, however, is not restricted to the above issues. Study of functions of different organs and systems made it possible to simulate these functions by means of the mechanical apparatus and devices. In this way an artificial kidney (apparatus for haemodialysis) was developed. Study of the physiology of the cardiac rhythm enabled the creation of the apparatus for heart electrostimulation which ensures normal cardiac performance and makes it possible for patients with serious cardiac abnormalities to resume their former work. The artificial heart and extracorporeal circulation apparatus (heart-lung machine) have been developed which enable the patient's heart to be switched off during the intricate cardiac surgery. There are also defibrillators which can restore the normal cardiac activity in cases of fatal derangement of the contractile myocardial function.

As a result of the research into the physiology of respiration, an apparatus for controlled artificial respiration (iron lungs) has been developed as well as various devices which provide for a possibility to switch off a patient's respiration for the operation period or to maintain life for years in subjects with lesions of the respiratory centre. A knowledge of the physiological patterns of gas exchange and transport provided conditions for creating hyperbaric oxygenation equipment, which is to be used in fatal diseases of the blood, respiratory, and cardiovascular systems. The techniques of a number of the most complicated neurosurgical operations have been elaborated on the basis of the laws governing the physiology of the brain. For instance, electrodes implanted into the cochlea of a deaf subject can perceive electric impulses coming from the artificial sound receptacles due to

which the hearing function can be restored to a certain degree. These are only a few examples illustrating the application of the laws of physiology to clinical practice, although the importance of physiology goes tar beyond the limits of clinical medicine.

THE ROLE OF PHYSIOLOGY IN ENSURING THE MAN'S VITAL ACTIVITY UNDER DIFFERENT CONDITIONS

Knowledge of physiology is required for the scientific substantiation of and providing conditions for a healthy way of life and prevention of disease. It is also indispensable for a scientific labour organization and the elaboration of various regimens for individual training of athletes and proper determination of sports exertions. This does not refer to sports alone. The laws of physiology help determine and provide for the conditions necessary for the life and work of people under extreme conditions - in the outer space and oceanic depths, in the North or South Pole, at high altitudes, in tundra or taiga, in exposure to extremely high or low temperatures, during movement to different time zones, and in various climatic conditions.

PHYSIOLOGY AND TECHNOLOGY

Knowledge of the laws of physiology was necessary not only for the scientific organization and raising of labour productivity. As is known, during billions of years of its evolution nature has achieved the highest perfection in designing and controlling the living organism's functions. Application of the principles, methods, and techniques operating in the organism to technology opened up new perspectives for a technological progress. In this way a new science, which is known as bionics, was born at the junction of physiology and technology. This science is concerned with the application of data about functioning of biological systems to the solution of engineering problems.

DEVELOPMENT OF METHODS OF PHYSIOLOGICAL STUDIES

Physiology emerged as an experimental science. It receives all the data by way of a direct investigation of the vital activity of the human and animal organisms. The forefather of the experimental physiology was a celebrated English physician William Harvey.

According to Pavlov, three hundred years ago, amids the deep darkness and an unimaginable confusion that reigned in the concepts on the vital activity of animal and human organisms, though these concepts were illuminated by the indisputable authority of the classical scientific heritage, William Harvey guessed at one of the most essential functions of the organism - circulation of the blood - thus laying down the foundation for a new part of accurate human knowledge, i.e. animal physiology.

But despite Harvey's discovery of the blood circulation, physiology developed slowly during the next two hundred years. Only a few basic works belonging to the 17-18th centuries can be mentioned. These were the discovery of the capillaries by Malpighi, the Descartes' formulation of the principle of reflex activity of the nervous system, the work of Heeth devoted to measurement of blood pressure, formulation of the law of conservation of matter by Lomonosov, the discovery of oxygen by Priestley and of the existence of a strict analogy between the processes of combustion and gas exchange by Lavoisier, Galvani's discovery of animal electricity, i.e. the capacity of living tissues for generating electric potentials, and some others.

Observation as a method of physiological study. The low level of production and stagnation in natural sciences as well as the difficulties encountered in the study of physiological phenomena by way of their ordinary observation are the causes of a rather slow development of experimental physiology during the two centuries after the discoveries made by W. Harvey. This methodological approach was and still remains the cause of numerous subjective errors. It is not an easy task to observe many complex processes and phenomena which develop and change incessantly during an experiment and to draw conclusions on their nature. Nor is it easy to establish an interconnection between them and other processes left unnoticed,

which is required for their analysis. Hence, ordinary observation is a source of subjective errors since it ensures only a qualitative assessment of physiological phenomena making their quantitative evaluation impossible.

Karl Ludwig's invention of the kymograph and introduction of the method for graphic recording of blood pressure into clinical practice (1843) was an important step in the development of experimental physiology.

Graphic recording of physiological processes. The method of graphic recording of physiological processes was a new step in the development of physiology. It enabled an objective registration of the process under study with a minimum of subjective errors. The experiment and analysis of the studied phenomenon could be performed in two stages. The task set before the experimenter was to obtain a high quality recording (curves) during the experiment; the obtained data could be analysed afterwards without drawing off the experimenter's attention during the experiment. This method enabled simultaneous recording of several (theoretically infinite number) of physiological processes.

Shortly afterwards, medical practice was enriched with a method for recording heart and muscle contractions (Engelmann). The device for pneumatic recording (Marey's capsule) of certain physiological processes occurring in the body was introduced by Marey. It became possible to record, and sometimes at a considerable distance from the object, such processes as respiratory movements in the thorax and abdominal cavity, peristalsis, and changes in the gastric and intestinal tone, etc. A method for measuring blood pressure in the arteries (Mosso's sphygmomanometer) was introduced into clinical practice as well as the method for measuring the volume of different internal organs (oncometry), etc.

Studies of the bioelectric phenomena. The discovery of the animal electricity' heralded quite an important trend in the development of physiology. The classical 'second experiment' suggested by Luigi Galvani demonstrated that living tissues were a source of electric potentials capable of exerting an influence on the nerves and muscles of other organisms and cause muscle contraction. Since that time and during nearly one hundred years the neuromuscular preparation of a frog served as

the sole indicator of the bioelectric potentials generated by the living tissues. With its help the potentials generated by the heart during its activity were discovered (Kelliker and Mtiller) as well as the fact that muscular contraction requires the incessant generation of electric potentials (Matteucci's experiment). It became clear that bioelectric potentials are not the chance or side phenomena in the living tissue activity but are signals for transmitting commands through the central nervous system to muscles and other organs thus bringing about interactions between living tissues using the “electric language”.

This language was comprehended much later, after the physical devices that could intercept the bioelectric phenomena were developed. Ordinary telephone was one of such devices. Using telephone, the Russian physiologist Vvedensky (Wedensky) discovered a number of most essential properties of nerves and muscles. Telephone proved to be helpful in 'listening' to bioelectric potentials, i.e. they were studied by way of observation. The next step was development of the method for objective graphic recording of bioelectric phenomena. The Dutch physiologist Einthoven invented a string galvanometer, an apparatus for recording electric potentials arising during cardiac performance on photographic paper. The recording is known as electrocardiogram (ECG). Samoilov, the disciple of Sechenov and Pavlov, was the first to introduce this method in this country (he worked for some time in Einthoven's laboratory in Leyden).

In actual fact, electrocardiography was soon transferred from physiological laboratories into clinical practice as a fairly reliable method for studying the heart, and today millions of patients owe their lives to this method.

Use of electronic amplifiers enabled subsequent development of compact electrocardiographs to record ECG in austronauts by means of telemetry or in athletes during competitions as well as in sick people living in remote areas from where ECG is transmitted by telephone to large cardiological centres for comprehensive analysis.

The objective graphic recording of bioelectric potentials formed the basis for emergence of the most essential branch of physiology - electrophysiology. The

English physiologist Adrian suggested the technique by which electron amplifiers may be applied for recording the bioelectric potentials. A Soviet scientist Pravdich-Neminsky was the first to record the biocurrents of the brain and to obtain electroencephalogram (EEG). This method was later improved by a Jena researcher Berger. Today, electroencephalography as well as the graphic recording of electric potentials of muscles (electromyography), nerves, and other excitable tissues has found wide application in clinical practice. This enabled the most accurate assessment of a functional status of a given organ or system. All these methods were highly significant for the science of physiology itself as they made it possible to decipher the functional and structural mechanisms underlying the nervous system activity, organ and tissue functioning, and regulation of physiological processes.

Invention of microelectrodes, which are small-calibre electrodes with the diameter at their tip equal to a fraction of a micron, was an important step in electrophysiology development. These electrodes can be introduced directly into the cell with the help of special devices or micromanipulators to record the bioelectric potentials within the cell. They help deciphering the mechanisms by which bioelectricity is generated, i.e. the processes occurring in the cell membrane. The cell membrane is the most important cellular formation which mediates the interaction between cells in the organism and between separate cellular elements. A science dealing with the biological membrane function (membranology) became an important part of physiology.

Methods for electric stimulation of organs and tissues. Living organs and tissues are capable of responding to any stimuli - thermal, mechanical, chemical, and others. By its nature electric stimulation appears to be closest to a 'natural language by which living systems exchange information. A Berlin physiologist Du Bois-Reymond was the founder of the well-known inductorium (induction coil) for electric stimulation of the living tissues with pulsing electric current.

At present, electron stimulators are used to obtain electric impulses of any form, frequency, and power. Electric stimulation has become an important method for

studying organ and tissue functioning and is widely used in clinical practice. Various designs of electric stimulators have been developed for implanting them into the body. Electric stimulation of the heart is now a reliable method for restoration of the normal rhythm and function of this vitally important organ owing to which hundreds of thousands of sick people were able to resume their work.

Electrostimulation of the skeletal muscles is being effectively used, and methods for electric stimulation of cerebral areas by means of implanted electrodes are being developed. The electrodes are introduced with the aid of special stereotaxic devices into the strictly defined nerve centres (with a precision reaching fractions of a millimetre). On being introduced into clinical practice, this technique provided cure of thousands of seriously ill patients with nervous disease. In addition, a lot of essential data on the mechanisms of the human brain activity have been obtained using this method (Bekhtereva).

This information can give an idea about certain methods of physiological studies and may serve an illustration to the importance of physiology for clinical practice.

Physiology makes wide use of the chemical methods along with the recording of electric potentials, temperature, pressure, mechanical movements, and other physical processes as well as the results of their influences on the body.

Chemical methods in physiology. The language of physical interactions by means of electric signals is not universal to the organism. The most common are chemical interactions, i.e. the chain of chemical processes taking place in living tissues. Hence the emergence of a new branch of chemistry, physiological chemistry, which deals with the study of these processes and has now turned into an independent science, biochemistry. The biochemical data reveal the molecular mechanisms of physiological processes. Experimental physiology widely applies chemical methods of study as well as those methods that have emerged at the junction of chemistry, physics, and biology. These, in turn, gave rise to new scientific research areas such as, for example, biophysics that deals with the physical aspects of physiological phenomena.

Physiology widely uses the method of labelled atoms. Other techniques borrowed from the exact sciences are also applied; they, indeed, provide information which is indispensable when certain mechanisms of physiological processes are to be analysed.

Electric recording of non-electrical values. Significant progress in physiology of today is associated with application of radioelectronic devices and apparatus. Among them are transducers, i.e. the devices that translate various nonelectrical phenomena and values (movement, pressure, temperature, concentration of various substances, ions, etc.) into electric potentials which are then intensified by electron amplifiers and recorded by oscillographs. A large number of various types of such recording devices have been elaborated by which many physiological processes can be recorded. In some of them use is made of an additional influence exerted on the body (ultrasound, electromagnetic waves, high-frequency electronic oscillations, etc.). Changes in the parameters of the influence altering certain physiological functions are recorded. The advantage of these devices lies in the fact that the pickup can be applied not to the organ being examined but to the body surface. Waves and oscillations penetrate the body, exert an influence on an organ functioning, and are then registered by the pickup. This principle is utilized in ultrasound flowmeters to determine the blood flow velocity in vessels and in rheographs and rheoplethysmographs which register changes in perfusion of different body parts. Another advantage is that the body can be examined at any time without a preliminary surgical intervention; therefore, these techniques are harmless and atraumatic. Most of the modern methods for physiological clinical examination are based on these principles. The pioneer in application of radioelectronic technology in the USSR was Academician V. V. Parin.

A significant advantage of these recording means consists in the fact that the studied physiological processes are converted by transducers into electric oscillations which can be amplified and transmitted by wires or radio to any distance from the examined object. The telemetry techniques provide for a laboratory examination of body functions in astronauts working in the outer space,

in pilots during flight, in athletes during competitions, or in workers during their working activity.

A more detailed analysis of physiological processes gives rise to the need for a synthesis by which separate elements can be united to create a whole picture of the phenomenon.

The aim of physiology is to provide a deeper analysis and synthesis of information thus treating the organism as an integral system.

The laws of physiology enable one to understand the reaction of the body as a whole system and of all its subsystems functioning under certain conditions and in the presence of a definite influence. Therefore, any method for exerting an influence on the body has to be tested comprehensively in physiological experiments prior to its introduction into clinical practice.

The method of acute experiment. A scientific progress is associated not only with the development of experimental technology and methods of study. It greatly depends on the evolution of physiological thinking and methodological approaches to the study of physiological phenomena. Beginning with its origin and till the eighties of the nineteenth century, physiology remained an analytical science, as it was concerned with the study of separate body organs and systems. Experiments on isolated organs, or the so-called acute experiments, were the main techniques of analytical physiology. To gain access to a definite internal organ or system vivisection had to be performed.

Animals were secured to a stand and complicated and painful operations were performed. It was a hard labour but, unfortunately, there was no other way to penetrate deep into the body. The matter concerned not only the moral aspect of the problem. Physical pain and unbearable sufferings experienced by the body crudely interfered with the normal course of physiological phenomena that is why the essence of the normal processes occurring under natural conditions could not be comprehended. Nor was the introduction of narcosis and other methods of analgesia into practice of essential help. In this way a vicious circle was created. Examination of a certain process or function of an internal organ or system

demanded deep penetration into the body, while attempts at such penetration distorted the course of vital processes for which the experiment was undertaken. In addition, study of isolated organs failed to give an idea about their actual functions in the intact organism.

The method of chronic experiment. The Russian science made its contribution to the development of physiology. Ivan Pavlov, one of the most talented Russian scientists, managed to find the way out of this dead alley. He was aware of the shortcomings of the analytical physiology and acute experiment and developed a method to look inside the body without interfering with its integrity. The method is known as the chronic experiment and is based on the 'physiological surgery'.

The method consisted essentially in the following. A complex operation was performed on an animal under conditions of anaesthesia, asepsis, and adherence to the rules of surgery in order to gain access to a certain internal organ. A 'window' was created in a hollow organ, a fistula made, or a glandular duct taken to the outside and sutured to the skin. The experiment was started much later after the wound had healed and the animal recovered. The applied fistula enabled a long-term study of the physiological processes under conditions of the animal's natural behaviour.

PHYSIOLOGY OF THE INTEGRAL ORGANISM

It is a common knowledge that the development of a science is closely related to the advances in technology.

A principally new science, or synthetic physiology, developed on the basis of the Pavlovian method of chronic experiment. This helped reveal the influence of the external environment on physiological processes and functional changes taking place in different organs and systems to provide for the organism's vital activity under various conditions.

With the development of new techniques for studying the vital processes it became possible to study functions of many internal organs not only in animals but in man as well without preliminary surgical manipulations. The physiological surgery

technique was replaced in some branches of physiology by modern methods of bloodless experiment. The main point, however, lies not in a certain technical method but in the methodology of physiological thinking. Pavlov created a new methodology, while physiology developed as a synthetic science with a systemic approach that is organically inherent in it.

The integral organism is intimately interrelated with its external environment and therefore, as Sechenov rightly pointed out, the scientific definition of the organism should include the environment that exerts its influence on it. Physiology of the integral organism is concerned with investigation of not only the internal mechanisms by which physiological processes are regulated but also the mechanisms that ensure the incessant interaction and inseparable unity of the organism and its environment.

Regulation of the vital processes as well as the interaction of the organism with its external environment are realized on the basis of the principles common to the machine and automated system control. These principles and laws make up the scope of cybernetics.

Cybernetics (Gk kybernetes steersman) is the science dealing with the general principles of control and communication in machines, mechanisms, and living organisms. It is known that processes of control are effected by way of signals which bear definite information. In the organism, the role of these signals is played by nerve impulses having electric nature and by various chemical substances.

The scope of cybernetics is the study of processes of perception, coding, processing, storage and reproduction of information. The organism has special devices and systems to fulfil these purposes (receptors, nerve fibres, nerve cells, etc.).

The models that reproduce certain functions of the nervous system have been created with the help of cybernetic devices. However, the cerebral activity as a whole cannot be simulated in this way and further research in this area is needed.

It was only thirty years ago that cybernetics began to be applied to physiology, but the mathematical and technical arsenal gained during this period ensured a considerable progress in the study and modelling of physiological processes.

Mathematics and computers in physiology. Synchronous recording of physiological processes makes possible their quantitative analysis and study of interactions between various phenomena; this involves exact mathematical methods whose application also heralded a new important stage in the development of physiology. Mathematization of investigations allows computers to be used in physiology, which not only accelerates the rate of information processing but makes it possible to perform this processing directly at the moment of the experiment, so that its course and the tasks of investigation can be changed in accordance with the results obtained.

So the loop of a spiral in the development of physiology has come to an end, as it were. The dawn of the development of this science was characterized by the method of observation which consisted in simultaneous investigation, analysis, and estimation of the results directly during the experiment. When graphic recording was invented, it became possible to perform the experimental stages separately, while the results could be treated and analysed after the experiment was completed. Radioelectronic and cybernetic devices enabled the investigator to perform analysis and processing of the information during the experiment but on a principally new basis: the interaction of many different physiological processes can be studied simultaneously and its result analysed quantitatively. In this way the possibility appeared to conduct the so-called controlled automatic experiment which makes use of computers not only to facilitate analysis of the results but to change the course and tasks of the experiment as well as the types of influence exerted on the organism, depending on the character of its response arising during the experiment. Physics, mathematics, cybernetics, and other exact sciences have equipped physiology with new technical means and provided physicians with a powerful arsenal of modern technological sources for the accurate functional assessment of the body status and for exerting an influence on the organism.

Mathematical simulation in physiology. The mathematical models of physiological processes have been created proceeding from the knowledge of physiological regularities and interrelationships existing between different physiological processes. Using these models, physiological processes are reproduced on computers which provide the diverse variants of reactions, i.e. their possible future changes produced by a certain influence on the body (drags, physical factors, or extreme environmental conditions). Currently the union of physiology and cybernetics has proved to be useful when serious surgical interventions are performed or other emergencies arise which require the most accurate evaluation of the current status of the essential physiological processes occurring in the body and foreseeing their possible changes. This approach considerably increases the reliability of the 'human factor' in complicated and important links of modern production.

Physiology of this century has achieved considerable successes in the area dealing with the discovery of the mechanisms underlying the vital processes and their control. In addition, it has made a breakthrough in the most intricate and mysterious branch of human knowledge - the area of psychological phenomena.

The physiological foundation of psyche, i.e. the human and animal higher nervous activity, has become one *of* the most essential objects of psychological investigation.

OBJECTIVE STUDY OF THE HIGHER NERVOUS ACTIVITY

In the course of millennia it was assumed that human behaviour is determined by the influence of a certain non-material essence or soul whose comprehension lies beyond the scope of the physiologist's perceptive power.

LM. Sechenov (Setchenow) was the first among the world physiologists who endeavoured to explain behaviour proceeding from the reflex principle, i.e. on the basis of the mechanisms of the nervous activity known in physiology. In his well-known book *Reflexes of the Brain* Sechenov showed that all the external manifestations of human psychic activity, no matter how complex they might seem, are reduced sooner or later solely to one thing - muscle movement. He

conceived that 'no matter whether a child smiles for a new toy, or Garibaldi laughs being kicked out for his immeasurable love for his native country, or Newton conceives the universal laws and writes them down on paper, or a girl trembles forseeing her first love encounter - *in* all these cases the final result of a mental activity is one thing - muscle movement'.

Sechenov consistently demonstrated that the development of the child's thinking is determined by the influences of the external environment which form various combinations and give rise to various associations. Our thinking is naturally shaped under the influence of the environmental conditions, while the brain is an organ that accumulates and reflects them. No matter how complex the manifestations of our psychic life seem to be, our inner psychological mentality is a natural result of conditions of education and environmental influences. Sechenov contended that the man's psychic content is mostly dependent on the factors of education and environment in a broad sense of the word, their ratio to the genetic factors being 999 to 1, respectively. Thus, *the principle of determinism* was applied for the first time to the most intricate field of life phenomena, to processes of man's spiritual life. According to Sechenov, there will be time when physiologists will be able to analyse the external manifestations of the cerebral activity with the same degree of precision as physicists analyse a musical chord. Sechenov's book brilliantly proved the materialistic concepts in the most intricate spheres of man's spiritual life.

Sechenov's endeavours to substantiate the mechanisms of the cerebral activity were purely theoretical, which called for the next step, for the experimental investigation of the physiological mechanisms underlying psychic activity and behaviour. This task was accomplished by Pavlov.

It was not by a mere chance that Pavlov became the successor of Sechenov's ideas and was the first to reveal the principal mysteries of the activity of the higher parts of the brain. It was the logic of his experimental physiological studies that brought him close to this subject. Studying the organism's vital activity in conditions of an animal's natural behaviour he noticed the important role of *psychic factors* that

influence all physiological processes. He was very keen to observe the fact that gastric juice and other digestive juices are secreted not at the moment of feeding an animal but long before, when the animal sees the food and hears the steps of the laboratory worker who usually feeds the animal. Pavlov also understood the role of appetite, the unbearable craving for food, which is as a powerful juice-secreting agent as the food itself. Factors such as appetite, craving, mood, emotions, and feelings are all the psychic phenomena which had not been the subject of study by the pre-Pavlovian physiologists. Pavlov believed that the physiologist could not ignore these phenomena since they actively interfere in the course of physiological processes and change their character. This is the subject matter of physiology. But how to study them? Before Pavlov *it* was zoopsychology that studied this subject.

Directing his attention to this science, Pavlov had to give up the firm basis of physiological facts to find himself in the area of the groundless and fruitless conjectures related to the animal's apparent psychic state. While human behaviour can be explained by the methods used in psychology and the man can describe his feelings, emotions, etc., zoophysiology used the data obtained during examinations of humans and applied to animals. Zoophysiologists also speculated about 'feelings', 'mood', 'emotions', 'wishes' and the like in the context of animal behaviour having no possibility to check if that was so or not. It was for the first time in Pavlov's laboratories that the number of opinions on the mechanisms of one and the same facts equalled the number of people dealing with these facts. Each fact was interpreted in a different way and there was no possibility to check the correctness of any of these interpretations. Since all these conjectures were non-sensical, Pavlov made a decisive and truly revolutionary step, i.e. he began *to study objectively animal behaviour* giving up all hopes to puzzle out the inner psychic state of animals. He compared the effects exerted on the body and the body response to them. This objective method of study helped reveal the regularities underlying the behavioural responses of an organism.

The method for the objective study of behavioural reactions formed the basis for a new science - *physiology of the higher nervous activity* with its thorough

knowledge of the processes which arise in the nervous system in response to environmental effects. This science has greatly contributed to the understanding of the mechanisms of human psychic activity.

Physiology of the higher nervous activity, the science that was created by Pavlov, became the *natural-scientific basis of psychology and the theory of reflection.* It is essential for understanding philosophy, medicine, and educational sciences, which deal with the study of psycho-emotional and intellectual powers of humans.

The importance of physiology of the higher nervous activity to medicine. Pavlov's theory of the higher nervous activity is of great practical importance. It is common knowledge that sick people are cured not only by means of drugs, scalpel, or some therapeutic procedure but also by the *word of a doctor*, by their confidence in the doctor's skill, and by their strong desire to be cured. All these facts were known in the time of Hippocrates and Avicenna. During the millennia, however, they were ascribed to the existence of a 'mighty soul given by God' that dominated over the 'burden body'. Pavlov's theory unravelled these mysteries and made it clear that the effect produced by magic things, by a sorcerer or by his oath which seemed magic are simply an example of the influence exerted by the higher nervous activity on the internal organs and on the regulation of all vital processes. The character of this influence is determined by the effect of the external environment on the body, the most important among which are *social conditions*, in particular human communication by means of thoughts expressed in words. Pavlov was the first in the history of science who demonstrated that the power of words consisted in that the words and speech, which regularly change behaviour and psychic status, are a special signalling system inherent only in man. The Pavlovian doctrine swept out idealism from the final seemingly inaccessible shelter, from the concept of a vital force given by God.

GENERAL PHYSIOLOGY OF EXCITABLE TISSUES

Each cell of the body, which consists of a hundred billion cells, has a very intricate structure and is capable of self-organization and interaction with other cells. Each

cell can accomplish so many actions and process so much information that they outnumber many times all those processes which are performed at a modern large production plant. Nevertheless, a cell is only one of the comparatively elementary subsystems in the complex hierarchy of systems that form a living organism.

All these systems are distinguished by a high orderliness. The normal functional structure of each system and the normal existence of each element of the system (including each cell) are determined by the continuous exchange of information between the elements (and between cells).

Information is transmitted by way of a direct (contact) interaction between cells, by transport of substances with tissue fluid, lymph, and blood (humoral link), and through transmission of bioelectric potentials from one cell to another which is the quickest way of information transmission in the body.

The multicellular organisms have developed a special system responsible for perception, transmission, storage, processing, and reproduction of information which is encoded in electric signals. This is the nervous system which has achieved its highest perfection in man. The nature of the bioelectric phenomena, or signals by which the nervous system transmits information, can be understood if we examine certain aspects of general physiology of the so-called *excitable tissues*, which include the nervous, muscular, and glandular tissues.

THE MECHANISMS OF REGULATION OF PHYSIOLOGICAL PROCESSES

The human organism, according to Pavlov, is a system (crudely speaking, a machine) which possesses the unique property of the highest degree of self-regulation. Proceeding from this, the same method is used both for the study of the human and any other system. It comprises the separation of the whole into the composite elements, study of the importance of each element, the interconnection of the elements and their interaction with the environment, which ultimately leads to the comprehension of the general principles of the work and control of the system. What is meant here is the concept of the *system approach.*

The *system approach* is the methodology of the scientific cognition which is based on the consideration of objects as systems. It binds a researcher to discover the integrity of an object and the variegated types of connections in it so as to create the general concept of the system. Objects of a high degree of complexity, such as*the human organism, have a multilevel organization in which systems of higher complexity are composed of more simple systems so that the hierarchical gradation of subsystems is formed. The elements in a system of any level intercommunicate by way of transmission of information. In animal and human organisms information is coded in the definite structure of biological molecules and in the definite pattern of nerve impulses (frequency, package sets, intervals between the packages, a certain temporal relationship of impulses and their packages in various nerve fibres, etc.).

Transmission of this information is a means by which regulation of processes is accomplished, i.e. the control of physiological functions, the activity of cells, tissues, organs, and systems, the organising behaviour, and the realization of the interaction between the organism and its environment.

The nervous system is the main regulatory (controlling) mechanism in the organism of higher animals and man, and reflex is the main mechanism underlying its activity.

Any reaction of the organism to the external world mediated by the CNS is called a reflex (L *reflectere* to bend back). The morphological substrate of such reactions is the reflex arc that comprises five links: (1) the receptor, a special device that perceives a definite type of external or internal environmental influences; (2) the afferent (sensory) neuron (or group of neurons) that transmits a signal arising in the receptor to the nerve centre; (3) the internuncial neuron (or interneuron) which is the central part of the reflex arc (or the nerve centre) of a given reflex; (4) the efferent (motor) neuron whose axon transmits a signal to the effector; (5) the effector which is a striated or smooth muscle or a gland which realizes the corresponding activity.

Any effector is therefore connected with a corresponding receptor by the elements of the reflex arc and is activated on stimulation of a given receptor. Spreading of excitation (signal), which emerges on stimulation of the receptor, along the reflex arc gives rise to the organising reaction.

The concept of the reflex was introduced in the first half of the seventeenth century by the French scientist Ren6 Descartes. It played the most essential role in the development of physiology as it unraveled the cause and the mechanisms of the organism's reactions and demonstrated the principle of determinism on which they are based (the principle of determinism is universal for both the inanimate and living nature, for the cause-effect relationship). In this way, an important contribution was made to the development of the materialistic notions concerning the mechanisms of the organism's reaction.

Since Descartes' times these reactions used to be considered machine-like, providing an automatic organism's response to stimulation of the receptor. However, such automatic reactions may take place only during the emergence of simple reflexes involving a limited number of the CNS links.

The organism's reflex reactions are commonly more complicated and involve the numerous CNS *links* (levels). *Reflexes in* this case are not reduced to simple unambiguous reactions but are links in the complex process of control of motor functions (behaviour) or the activity of the internal organs.

separate machine-like reflex responses, is more intricate and is marked by certain general features and regularities inherent in the animal or human organism.

These general features is the scope of the science known as *cybernetics* which deals with the general features and laws of control realized on the basis of perception, storage, transmission, and processing of information whatever the physical nature of the object or system concerned.

Study of the laws of cybernetics and comprehension of their meaning is important for grasping the essence of the regulation of physiological functions and their modelling, whether mathematical or experimental, for their automatic control and

for the interference in deranged physiological processes in order to bring them to norm.

It is known that processes of control and automatic regulation were utilized in technology long before they were discovered *in the* organism and prior to the formulation of the laws of cybernetics.

According to Sechenov, machines possess regulators which are substitutes for the operator's hand; they are set in expedient work performance by themselves, as it were. In actual fact, they work under the influence of the changing conditions in machine operation. Such is, for example, a safety valve in the steam engine of Watt. As soon as the steam tension in the broiler grows above the given level, the valve itself increases the steam outlet and vice versa. A multitude of devices of this type are known and all are called automatic regulators. *In* the animal body, as in the self-operating machines, *the regulators apparently can be only automatic, i.e. they are put into action under the influence of changing conditions in the state or activity of the machine (organism) and they can develop an activity by which all defects are eliminated.* This concept was formulated by Sechenov in 1897; it demonstrates the principles of cybernetics as applied to the mechanisms of the organism's self-regulation.

Thus, Sechenov formulated the principle of the *negative feedback* on which the processes of self-regulation in the machine and living organism are based.

Many physiological processes are regulated according to this principle. Claude Bernard, the French physiologist and pathologist, was the first to discover the importance of the constancy of the internal environment (internal milieu) for the organism's life. He showed, for example, that any deviation from norm of blood sugar level brought into play the processes which balanced these deviations to maintain the constant level of this parameter in the organism. This principle underlies regulation of the constancy of body temperature in the homoiothermic animals as well as the other parameters of the internal milieu.

The German scientist Karl Ludwig and the Russian physiologist F. Zion discovered the same mechanism (operating by the negative feedback principle)

which regulates a constant level of arterial pressure. The sensory (depressor) nerve endings located in the aortic arch send powerful signals to the CNS when the blood pressure rises in this vessel. They give rise to a reflex slowing down of the heart beat and causes dilatation of arterioles, which leads to a fall in arterial pressure, i.e. to its return to the initial level. Later a great number of such regulatory mechanisms were discovered in the organism. The importance of the negative feedbacks in the regulation of movements, i.e. the signals arriving from the working muscles, was emphasized by Sechenov. *The positive feedback* mechanism was also discovered in a number of physiological processes which enables the arising process to be intensified and maintained by itself.

The feedback is a process whereby a part of the output of a system is returned to the input. It detects some deviations already aroused in the state of a system. The regulatory mechanisms based on it operate according to the *mismatching* principle. They are switched in the activity when the state of a system deviates from a given magnitude, i.e. when there is mismatching between a given (necessary) and actually existing magnitude. The mechanisms operating according to this principle are widely distributed in the organism. The general principle of their operation was given by Anokhin in his scheme of a functional system (Fig. 237, Vol. II). This scheme is, however, not universal since the work of the regulatory mechanisms existing in the organism is based on some other principle. The signal that triggers their activity is the deviation from a given magnitude not at the output but at the input of the system, i.e. action on the system of a stimuli differing from the given parameters. In this case, the regulatory reactions are based on another principle, i.e. on the work of a regulator according to the *disturbance* principle. At the output of the system there are devices which detect the magnitude of the incoming signal disturbing the system's state. If this magnitude exceeds the permissible one so as to give rise to an untoward deviations in the system's state, then commands emerge which neutralize the action of these signals and maintain a stable state of this system. Here arises prevention of the possibility of such disturbances rather than the recovery pf the already disturbed state of the system. (Both these principles of

the maintenance of the system's stability differ from each other as, for example, *differ* the means used for extinguishing fire that had already begun from those used for its prevention.)

The interplay of these two principles and of both regulatory mechanisms which function at the output and input of a system is found in any physiological regulatory, defence, and compensatory reactions. For example, when the eye is exposed to the damaging action of a stream of dusty air both mechanisms start operating. The blinking reflex prevents dirt from getting into the eye by closing it (this mechanism operates at the system's input according to disturbance), while the reflexogenic lacrimation and washing of the sclera and cornea by tears removes dirt (the mechanism operating at the system's output) according to mismatching principle. The combination of the action of these two mechanisms operating according to these different principles can be encountered in any homeostatic reaction.

Any regulatory reaction requires an information on the status of the system, the strength of the arriving signals, and on the emerging changes. An apparatus is needed for comparing the parameters of these changes or those of the arriving signals with the *magnitude* of the parameters which are normal for a given system. In addition, an apparatus is necesary to issue commands which would prevent these changes. The action of these commands is accomplished in two ways: (a) normalization of the already developed deviations (the mechanisms operating according to the mismatching principle) and (b) prevention of unfavourable effects of the input (disturbing) signal by decreasing its strength, preventing its action, or reducing the system's sensitivity to the given disturbing effect (the mechanism operating according to the disturbance principle). The regulatory responses in the organism are effected by the nervous system.

GENERAL PHYSIOLOGY OF THE CENTRAL NERVOUS SYSTEM

The central *nervous system* coordinates the *activity of* all organs and systems of the body, ensures effective adjustment of the organism to the environmental changes, and is involved in the formation of a directed behaviour. These complicated and

vitally important tasks are performed by the *nerve cells* or *neurons,* which have a special function of the *reception, processing, storage,* and *transmission* of information and are united into specifically organized neuronal chains making up various *functional systems of the brain.*

Nerve cells are interconnected by *synoptic contacts* which ensure the transmission of electrical *signals from* one neuron *to* another.

The number of nerve elements, being very limited in the primitive organisms, has reached many billions *in* the primates and man during evolution. The number of synaptic interneuronal contacts approaches an enormous figure of 10^{15}. The complexity of the CNS organization is manifested in the structural and functional variation of neurons in the different brain divisions. Nevertheless, studies of the various brain parts and cells of the nervous system of animals at different evolution levels have revealed certain general patterns which determine the course of the major nervous processes, *excitation* and *inhibition,* in the CNS neurons and synapses. The brain activity can be analysed by outlining the general fundamental principles which underly the functioning of the neurons and synapses.

Mental Disorders and Their Treatment

Introduction

Mental disorders, in particular their consequences and their treatment, are of more concern and receive more attention now than in the past. Mental disorders have become a more prominent subject of attention for several reasons. They have always been common, but, with the eradication or successful treatment of many of the serious physical illnesses that formerly afflicted humans, mental illness has become a more noticeable cause of suffering and accounts for a higher proportion of those disabled by disease. Moreover, the public has come to expect the medical profession to help it obtain an improved quality of life in its mental as well as physical functioning. And indeed, there has been a proliferation of both pharmacological and psychotherapeutic treatments in psychiatry in this regard, many of which have proved effective. The transfer of many psychiatric patients, some still showing conspicuous symptoms, from mental hospitals into the community has also increased the public's awareness of the importance and prevalence of mental illness.

There is no simple definition of mental disorder that is universally satisfactory. This is partly because mental states or behaviour that are viewed as abnormal or pathological in one culture may be regarded as normal or acceptable in another, and in any case it is difficult to draw a line clearly demarcating healthy from pathological mental functioning.

A narrow definition of mental illness would insist upon the presence of organic disease of the brain, either structural or biochemical; however, this condition does not pertain, as far as is known, to the majority of mental disorders. An overly broad definition would define mental illness as simply being the lack or absence of mental health - that is to say, a condition of mental well-being, balance, and resilience in which the individual can successfully work and function and in which he can both withstand and learn to cope with the conflicts and stresses encountered in life. A more generally useful definition than either of the above is that a mental

disorder is an illness with significant psychological or behavioral manifestations that occurs in an individual and that is associated either with a painful or distressing symptom, with impairment in one or more important areas of functioning, or with both. The mental disorder may be due to either a psychological, social, biochemical, or genetic dysfunction or disturbance in the individual.

A mental illness can have an effect on every aspect of a person's life, including thinking, feeling, mood, and outlook and such areas of external activity as family and marital life, sexual activity, work, recreation, and management of material affairs. Most mental disorders negatively affect how an individual feels about himself and impair his capacity for participating in mutually rewarding relationships.

Psychopathology is the systematic study of the significant causes, processes, and symptomatic manifestations of mental disorders. The meticulous study, observation, and enquiry that characterize the discipline of psychopathology are in turn the basis for the practice of psychiatry - i.e., the science and practice of treating mental disorders, as well as dealing with their diagnosis and prevention.

Psychiatry and its related disciplines in turn embrace a wide spectrum of techniques and approaches for treating mental illnesses. These include the use of psychoactive drugs to correct biochemical imbalances in the brain or otherwise to relieve depression, anxiety, and other painful emotional states.

Another important group of treatments are the psychotherapies, which seek to treat mental disorders by psychological means and which involve verbal communication between the patient and a trained person in the context of a therapeutic interpersonal relationship between them. An important variant of this latter mode of treatment is behavioral therapy, which concentrates on changing or modifying observable pathological behaviours by the use of conditioning and other experimentally derived principles of learning.

Types and causes of mental disorders

CLASSIFICATION AND EPIDEMIOLOGY

Psychiatric classification attempts to bring order to the enormous diversity of mental symptoms, syndromes, and illnesses that are encountered in clinical practice. Epidemiology is the measurement of the prevalence, or frequency of occurrence, of these psychiatric disorders in different human populations.

Classification.

Diagnosis is the process of identifying an illness by studying its signs and symptoms and by considering the patient's history. Much of this information is gathered by the psychiatrist during his initial interviews with the patient, who describes his main complaints and symptoms and any past ones and briefly gives his personal history and current situation. The psychiatrist may administer any of several psychological tests to the patient and may supplement these with a physical and a neurological examination. These data, along with the psychiatrist's own observations of the patient and of the patient's interaction with him, form the basis for a preliminary diagnostic assessment. For the psychiatrist, diagnosis involves finding the most prominent or significant symptoms, upon which the patient's disorder can be assigned to a category as a first stage toward rational treatment. This is as essential in psychiatry as in the rest of medicine.

Classificatory systems in psychiatry aim to distinguish groups of patients who share the same or related clinical symptoms in order to provide an appropriate therapy and accurately predict the prospects of recovery for any individual member of that group. Thus, the diagnosis of, for example, depressive illness having been made, it becomes logical to consider antidepressant drugs when preparing a course of treatment.

The diagnostic terms of psychiatry have been introduced at various stages of the discipline's development and from very different theoretical standpoints. Sometimes two words with quite different derivations have come to mean almost the same thing, for example, dementia praecox and schizophrenia. Sometimes a word, such as hysteria, carries many different meanings depending on the psychiatrist's theoretical orientation.

Psychiatry is hampered by the fact that the cause of many mental illnesses is unknown, and so convenient diagnostic distinctions cannot be made among such illnesses as they can, for instance, in infectious medicine, where infection with a specific type of bacterium is a reliable indicator for a diagnosis of tuberculosis. But the greatest difficulties presented by mental disorders as far as classification and diagnosis are concerned is that the same symptoms are often found in patients with different or unrelated disorders, or a patient may show a mix of symptoms properly belonging to several different disorders. Thus, although the categories of mental illness are defined according to symptom patterns, course, and outcome, the illnesses of many patients constitute intermediate cases between such categories, and the categories themselves may not necessarily represent distinct disease entities and are often poorly defined.

The two most frequently used systems of psychiatric classification are the International Classification of Diseases produced by the World Health Organization and the Diagnostic and Statistical Manual of Mental Disorders (DSM-III) produced by the American Psychiatric Association. The ninth revision of the former, published in 1977, is widely used in western Europe and other parts of the world for epidemiological and administrative purposes. Its nomenclature is deliberately conservative in conception so that it can be used by clinicians and mental health care systems in different countries.

This article, however, will follow the third edition of the DSM-III, which was published in 1980 and revised in 1987. The DSM-III differs from the International Classification in its introduction of precisely described criteria for each diagnostic category; its categorizations are usually based upon the detailed description of symptoms.

The DSM-III has been widely used, especially in the United States, and its detailed descriptions of diagnostic criteria have been useful in eradicating the inconsistencies of earlier classifications. However, there are still some major problems in its everyday clinical use. Chief among these problems is the DSM-III's innovative and controversial abandonment of the general categories of psychosis

and neurosis in its classificatory scheme. These terms have been and still are widely used to distinguish between classes of mental disorders.

Psychoses are major mental illnesses that are characterized by severe symptoms such as delusions, hallucinations, disturbances of the thinking process, and defects of judgment and insight. Psychotic patients exhibit a disturbance or disorganization of thought, emotion, and behaviour so profound that they are often unable to function in everyday life and may be incapacitated or disabled. The psychotic patient is often unable to realize that his subjective perceptions and feelings do not correlate with objective reality, a phenomenon evinced by psychotics who do not know or will not believe that they are ill despite the distress they feel and their obvious confusion concerning the outside world. The psychoses are broadly divided into organic and functional psychoses. In organic psychoses the mental disturbance results from a physical defect of or damage to the brain. In functional psychoses, notably the schizophrenias and affective psychoses, no physical brain disease is evident upon clinical examination, but there is some research evidence pointing to an underlying organic abnormality.

Neuroses, or psychoneuroses, are less serious disorders in which a person may experience such negative feelings as anxiety or depression and his functioning may be significantly impaired, but his personality remains relatively intact, he maintains a capacity for recognizing and objectively evaluating reality, and he is basically able to function in everyday life. In contrast to the psychotic, the neurotic patient knows or can be made to realize that he is ill, and he usually wants to get well and return to a normal state. His chances for recovery are better than those of the psychotic patient. The symptoms of neurosis may sometimes resemble the coping mechanisms used in everyday life by most people, but in the neurotic person these defensive reactions are inappropriately severe or prolonged in response to an external stress. Anxiety disorders, phobic disorders, conversion disorders, obsessive-compulsive disorders, and depressive disorders have been traditionally classed as neuroses.

There are various other mental illnesses, such as personality disorders (or "character disorders"), that cannot be classed as either psychoses or neuroses.

Epidemiology.

Epidemiology is the study of the distribution of disease in different populations. Prevalence denotes the number of cases of a condition present at a particular time or over a specified period, while incidence denotes the number of new cases occurring in a defined time period. Epidemiology is also concerned with the social, economic, or other contexts in which mental illnesses arise.

The understanding of mental disorders is aided by knowledge of the rate and frequency with which they occur in different societies and cultures. Looking at the worldwide prevalence of mental disorders reveals many surprising findings. It is remarkable, for instance, how constant the rate for schizophrenia is; in widely different cultures there is generally a lifetime risk of developing the illness of just under 1 percent.

Gradual historical changes in the incidence and prevalence of particular disorders have often been described, but it is very difficult to obtain firm evidence that such changes have actually occurred. On the other hand, prevalence has been seen to increase for a few syndromes due to general changes in living conditions over time. For example, dementia inevitably develops in some 20 percent of those persons over age 80, so that with the increase in life expectancy common to developed countries the number of people with dementia is bound to increase. Other factors, such as the presence of small quantities of aluminum in drinking water, may also play a part in the increased prevalence of dementia. There also seems to be some evidence of an increased prevalence of affective disorders over the last century.

Several large-scale epidemiological studies have been conducted to determine the incidence and prevalence of mental disorders in the general population. Simple statistics based on those people actually under treatment for mental disorders cannot be relied upon in making such a determination, because the number of those

who have sought treatment is substantially smaller than the actual number of people afflicted with mental disorders, many of whom do not seek professional treatment. Moreover, surveys to determine incidence and prevalence depend for their statistics on the clinical judgment of the survey takers, which can always be fallible because there are no objective tests for the assessment of mental illness. Given such objections, one ambitious study conducted by the National Institutes of Mental Health in the United States examined thousands of persons in several American localities and yielded the following results concerning the prevalence of mental disorders in the general population. About 0.6 percent of those surveyed were found to be schizophrenic, 0.5 percent had a manic episode, 5 percent suffered from depression, 5 percent suffered from phobias or other anxiety disorders, and about 1 percent had obsessive-compulsive disorders.

There is a relatively strong epidemiological association between socioeconomic class and the occurrence of certain types of mental disorders and of general patterns of mental health. One study found that the lower the socioeconomic class, the greater the prevalence of psychotic disorders; schizophrenia was found to be 11 times more frequent among the lowest of the five classes surveyed (unskilled manual workers) than among the highest class (professionals). (Anxiety disorders were found to be more common among the middle class, however.) Two possible explanations for the elevated frequency of schizophrenia among the poor would be that schizophrenics "drift downward" to the lowest socioeconomic class because they are impaired by their illness, or alternatively that unfavourable sociocultural conditions create circumstances that help induce the illness.

The manifestation of particular psychiatric symptoms is sometimes closely associated with particular epochs or periods in life. The symptoms of infantile autism are usually evident by early childhood, for example. Childhood and adolescence may produce a variety of psychiatric symptoms peculiar to those periods of life. Anorexia nervosa, several types of schizophrenia, drug abuse, and manic-depressive psychosis often first appear during adolescence or in young adult life. Alcohol dependence and its consequences, paranoid schizophrenia, and

repeated attacks of depressive illness are more likely to occur in middle age. Involutional melancholia and presenile dementias typically occur in late middle age, while senile and arteriosclerotic dementias are characteristic of the elderly.

There are also marked sex differences in the incidence of certain types of mental illness. For instance, anorexia nervosa is 20 times more common in girls than boys; schizophrenia occurs more commonly in men than women, and they tend to develop the illness at a younger age; depressive illness is more common in women than men; and many sexual deviations occur almost exclusively in men.

THEORIES OF CAUSATION

Very often the etiology, or cause, of a particular type of mental disorder is unknown or is understood only to a very limited extent. The situation is complicated by the fact that a mental disorder such as schizophrenia may be caused by a combination and interaction of several factors, including a probable genetic predisposition to develop the disease, a postulated biochemical imbalance in the brain, and a cluster of stressful life events that help to precipitate the actual onset of the illness. The predominance of these and other factors probably varies from patient to patient in schizophrenia. A similarly complex interaction of constitutional, developmental, and social factors can influence the formation of neurotic disorders.

No single theory of causation can explain all mental disorders or even all those of a particular type, and, moreover, the same type of disorder may have different causes in different patients; e.g., an obsessive-compulsive disorder may have its origins in a biochemical imbalance, in an unconscious emotional conflict, in faulty learning processes, or in a combination of these. The fact that quite different therapeutic approaches can produce equal improvements in different patients with the same type of disorder underscores the complex and ambiguous nature of the causes of mental illness. The major theoretical and research approaches to the causation of mental disorders are treated below.

Organic and hereditary etiologies.

Organic explanations of mental illness have usually been genetic, biochemical, neuropathological, or a combination of these.

Genetics.

The study of the genetic causes of mental disorders involves the statistical analysis of the frequency of a particular disorder's occurrence among individuals who share related genes; i.e., family members and particularly twins. Family risk studies compare the observed frequency of occurrence of a mental illness in close relatives of the patient with its frequency in the general population. First degree relatives (parents, siblings, and children) share 50 percent of their genetic material with the patient, and higher rates of the illness in these relatives than expected indicate a possible genetic factor. In twin studies the frequency of occurrence of the illness in both members of pairs of identical (monozygous) twins is compared with its frequency in both members of a pair of fraternal (dizygous) twins. A higher concordance for disease among the identical than the fraternal twins suggests a genetic component. Further information on the relative importance of genetic and environmental factors accrues from comparing identical twins reared together with those reared apart. Adoption studies comparing adopted children whose biological parents had the illness with those whose parents did not can also be useful in separating biological from environmental influences.

Such studies have pointed up a clear role for genetic factors in the causation of schizophrenia. When one parent is found to have the disorder, the probability that his children will develop schizophrenia is at least 10 times higher (about a 12-percent risk probability) than it is for children in the general population (about a 1-percent risk probability). If both parents are schizophrenic, their children stand anywhere from a 35- to 65-percent probability of becoming schizophrenic. If one member of a pair of fraternal twins develops schizophrenia, there is about a 10-percent chance that the other twin will also develop the disorder. If one member of a pair of identical twins has schizophrenia, the other identical twin has at least a 40-50 percent chance of developing the disease.

Genetic factors seem to play a less significant role in the causation of other psychotic disorders and in personality disorders, and they seem to be even less of a factor in the causation of the neuroses.

Biochemistry.

If a mental disease is caused by a biochemical abnormality, investigation of the brain at the site where the biochemical imbalance occurs should show neurochemical differences from normal. In practice such a simplistic approach is fraught with practical, methodological, and ethical difficulties. The living human brain is not readily accessible to direct investigation, and the dead one undergoes chemical change; moreover, findings of abnormalities in cerebrospinal fluid, blood, or urine may have no relevance to the question of a presumed biochemical imbalance in the brain. Human mental illnesses cannot be adequately studied using animals as analogues, because most mental disorders either do not occur or are not recognizable in animals. Even when biochemical abnormalities have been found in mental patients, it is difficult to know whether such abnormalities are the cause or the result of the illness, or of its treatment, or of other consequences. Despite these problems, progress has been made in unraveling the biochemistry of affective disorders, schizophrenia, and some of the dementias.

Certain drugs have been demonstrated to have beneficial effects upon mental illnesses. Antidepressant, antipsychotic, and antianxiety drugs are thought to achieve their therapeutic results by the selective inhibition or enhancement of the quantities, action, or breakdown of neurotransmitters in the brain. Neurotransmitters are a group of chemical agents that are released by neurons (nerve cells) to stimulate neighbouring neurons, thus allowing impulses to be passed from one cell to the next throughout the nervous system. Neurotransmitters play a key role in transmitting nerve impulses across the microscopic gap (synaptic cleft) that exists between neurons. The release of such neurotransmitters is stimulated by the electrical activity of the cell. Among the principal neurotransmitters are norepinephrine, dopamine, and serotonin. Some

neurotransmitters excite or activate neurons, while others act as inhibiting substances. Abnormally low or high concentrations of neurotransmitters at sites in the brain are thought to change the synaptic activities of neurons, thus ultimately leading to the disturbances of mood, emotion, or thought found in various mental disorders.

Neuropathology.

In the past the pathological study of the brain at post mortem revealed information upon which great advances in understanding the etiology of neurological and some mental disorders were based, leading to the German psychiatrist Wilhelm Griesinger's postulate: "all mental illness is disease of the brain." The application of the principles of pathology to general paralysis of the insane, one of the most common conditions found in mental hospitals in the late 19th century, resulted in the discovery that this was a form of neurosyphilis and was caused by infection with the spirochete bacterium Treponema pallidum.

The examination of the brains of patients with other forms of dementia has given useful information concerning other causes of this syndrome, for example, Alzheimer's disease and arteriosclerosis. The pinpointing of abnormalities of specific areas of the brain has aided understanding of some abnormal mental functions, such as disturbances of memory or speech disorders. However, no abnormal pathology has been demonstrated for most mental disorder.

Psychodynamic etiologies.

Up to the 1970s theories of the etiology of mental disorders, especially of neuroses and personality disorders, were dominated in the United States by Freudian psychoanalysis and the derivative theories of the post-Freudians. In western Europe the influence of Freudian theory upon psychiatric theory diminished after World War II.

Theories of personality development.

Freudian and other psychodynamic theories view neurotic symptoms as arising from intrapsychic conflict; that is, as being caused by conflicting motives, drives, impulses, and feelings held within various components of the mind. Central to psychoanalytic theory is the postulated existence of the unconscious, which is that part of the mind whose processes and functions are inaccessible to the individual's conscious awareness or scrutiny. One of the functions of the unconscious is thought to be that of a repository for traumatic memories, feelings, ideas, wishes, and drives that are threatening, abhorrent, anxiety-provoking, or socially or ethically unacceptable to the individual. These mental contents may at some time be pushed out of conscious awareness but remain actively held in the unconscious. This process is a defense mechanism for protecting the individual from the anxiety or other psychic pain associated with those contents and is known as repression. The repressed mental contents held in the unconscious retain much of the psychic energy or power that was originally attached to them, however, and they can continue to influence significantly the mental life of the individual even though (or because) he is no longer aware of them.

The natural tendency for repressed drives or feelings, according to this theory, is to break through into conscious awareness so that the individual can seek the gratification, fulfillment, or resolution of them. But this threatened release of forbidden impulses or memories provokes anxiety and is seen as threatening, and a variety of psychic processes known as defense mechanisms may then come into play to provide relief from the state of psychic conflict. Through reaction formation, projection, regression, sublimation, rationalization, and other defense mechanisms, some component of the unwelcome mental contents can emerge into consciousness in a disguised or attenuated form, thus providing partial relief to the individual. Later, perhaps in adult life, some event or situation in the person's life triggers the abnormal discharge of the dammed-up or strangulated emotional energy in the form of neurotic symptoms in a manner mediated by defense mechanisms. Such symptoms can form the basis of neurotic disorders such as

conversion and somatoform disorders, anxiety disorders, obsessional disorders, and depressive disorders. Since the symptoms represent a compromise within the mind between letting the repressed mental contents out and continuing to deny all conscious knowledge of them, the particular character and aspects of an individual's symptoms and neurotic concerns bear an inner meaning that symbolically represents the underlying intrapsychic conflict. Psychoanalysis and other dynamic therapies associate the patient's controlled and therapeutic recovery to conscious awareness of repressed mental conflicts, and his understanding of their influence on both his past history and present difficulties, with the relief of symptoms and improved mental functioning.

Freudian theory views childhood as the primary breeding ground of neurotic conflicts. This is because children are relatively helpless and are dependent on their parents for love, care, security, and support and because their psychosexual, aggressive, and other impulses are not yet integrated into a stable personality framework. Children are thus liable to emotional traumas, deprivations, and frustrations which they lack the resources to cope with and which can become grounds for intrapsychic conflicts that are not resolved but rather are merely held in abeyance through repression, producing insecurity, unease, or guilt and subtly influencing the individual's developing personality, interests, attitudes, and ability to cope with later stresses.

Non-Freudian psychodynamics.

Psychoanalytic theory's emphasis on the unconscious mind and its influence on human behaviour resulted in a proliferation of other, related theories of causation incorporating many basic psychoanalytic precepts. Most subsequent psychotherapies have stressed in their theories of causation aspects of earlier, maladaptive psychological development that had been missed or underemphasized by orthodox psychoanalysis, or they have incorporated insights taken from learning theory. The Swiss psychiatrist Carl Jung, for instance, concentrated on the individual's need for spiritual development and concluded that neurotic symptoms

could arise from a lack of self-fulfillment in this regard. The Austrian psychiatrist Alfred Adler emphasized the importance of feelings of inferiority and the unsatisfactory attempts to compensate for it as important causes of neurosis. Neo-Freudian authorities such as Harry Stack Sullivan, Karen Horney, and Erich Fromm modified Freudian theory by emphasizing social relationships and cultural and environmental factors as being important in the formation of mental disorders. Many other highly specific theories of causation have been developed by particular psychotherapies, and in general, psychiatric scrutiny has come to extend far beyond the confines of early psychosexual development that were originally posited by Freud as the prime ground for the causation of neuroses. More modern psychodynamic theories have moved away from the idea of explaining and treating neurosis on the basis of a defect in a single psychological system and have instead adopted a more complex notion of multiple causes, including emotional, psychosexual, social, cultural, and existential ones. A notable trend in the more recently developed psychotherapies has been the incorporation of approaches derived from theories of learning. Such psychotherapies pay special attention to acquired, faulty mental processes and maladaptive behavioral responses that act to sustain neurotic symptoms, and there has generally been increased interest in the patient's present circumstances and his learned responses to those conditions as a causative factor in mental illness. In this way, psychoanalytic theory and behavioral theory have tended somewhat to converge and intermingle in their views of disease causation.

Behavioral etiology.

Behavioral theories for the causation of mental disorders, especially neurotic symptoms, are based upon learning theory, which was in turn largely derived from the study of the behaviour of animals in laboratory settings. Most important theories in this area arose out of the work of the Russian physiologist Ivan Pavlov and such American psychologists as Edward L. Thorndike, Clark L. Hull, John B. Watson, Edward C. Tolman, and B.F. Skinner. In the classical Pavlovian model of

conditioning, an unconditioned stimulus is followed by an appropriate response; for example, food placed in a dog's mouth is followed by the dog salivating. If a bell is rung just before food is offered to a dog, eventually the dog will salivate at the sound of the bell only, even though no food is offered. Because the bell could not originally evoke salivation in the dog (and hence was a neutral stimulus) but came to evoke salivation because it was repeatedly paired with the offering of food, it is called a conditioned stimulus. The dog's salivation at the sound of the bell alone is called a conditioned response. If the conditioned stimulus (the bell) is no longer paired with the unconditioned stimulus (the food), extinction of the conditioned response gradually occurs (the dog ceases to salivate at the sound of the bell alone).

Behavioral theories for the causation of mental disorders rest largely upon the assumption that the symptoms or symptomatic behaviour found in persons with various neuroses (particularly phobias and other anxiety disorders) can be regarded as learned behaviours that have been built up into conditioned responses. In the case of phobias, for example, a person who has once been exposed to an inherently frightening situation afterward experiences anxiety even at neutral objects that were merely associated with that situation at the time but that should not reasonably produce anxiety; e.g., a child who has had a painful session with a dentist may subsequently have an unreasonable dread of men in white coats or of any kind of drill. The neutral object alone is enough to arouse anxiety, and the person's subsequent effort to avoid that object is a learned behavioral response that is self-reinforcing, since the person does indeed procure a reduction of his anxiety by avoiding the feared object and is thus likely to continue to avoid it in the future. But his fear of the object persists, since it is only by confronting the object that he can eventually lose his irrational, association-based fear of it.

Other etiologies.

Mental illness may be contagious in a psychological sense; that is, close contact with an individual who has symptoms may result in the transmission of those

symptoms to one or many others who were previously unaffected. This may occur either through the powerful influence of long-term cohabitation of one person with one other - a phenomenon known as folie à deux - or through the volatile collective emotions of a group - mass hysteria. Epidemic, communicated, or mass neurosis is particularly likely to spread through a closed community such as a boarding school. The transmission of symptoms occurs first to those who are psychologically vulnerable, and stops once the closed population is scattered.

Social values can sometimes determine or encourage the formation of particular syndromes. Prime examples of this are anorexia nervosa and bulimia nervosa, which predominantly affect young females in affluent Western societies. The value and attractiveness of physical slimness are communicated via the media and respected adults, and an eating disorder resulting in emaciation subsequently occurs in some neurotically susceptible individuals.

Another approach to the causation of mental disorders focuses on the effects and consequences of stress, which is a state of bodily or mental tension resulting from external factors such as marital conflicts, excessive work demands, or serious financial problems. Stress is known to cause psychosomatic illnesses, and an accumulation of stressful life events can help cause depression in psychologically vulnerable individuals.

MAJOR DIAGNOSTIC CATEGORIES

Organic mental disorders.

This category includes both those psychological or behavioral abnormalities that arise from structural disease of the brain and also those that arise from brain dysfunction caused by disease outside the brain. These conditions differ from those of other mental illnesses in that they have a definite and ascertainable cause - i.e., brain disease. Treatment, when possible, is aimed at both the symptoms and the underlying physical dysfunction in the brain.

There are several types of psychiatric syndromes that arise from organic brain disease, chief among them being dementia and delirium. Dementia is a gradual and progressive loss of such intellectual abilities as thinking, remembering, paying attention, judging, and perceiving, without an accompanying disturbance of consciousness. The syndrome may also be marked by the onset of personality changes. Dementia is usually a chronic condition and frequently worsens over the long term. Delirium is a diffuse or generalized intellectual impairment marked by a clouded or confused state of consciousness, an inability to attend to one's surroundings, difficulty in thinking coherently, a tendency to perceptual disturbances such as hallucinations, and difficulty in sleeping. Delirium is generally an acute condition and is not long-lasting. Other specific psychological impairments associated with organic brain disease are amnesia (a gross loss or disorder of recent memory and time-sense without other intellectual impairment), recurring or persistent hallucinations or delusions, or marked personality changes.

In the diagnosis of suspected organic disorders, a full history has first to be taken from the patient and his mental state must be examined in detail, with additional tests for particular functions added if necessary. A physical examination is also carried out with special attention to the central nervous system. In order to determine whether a metabolic or other biochemical imbalance is causing the condition, blood and urine tests, liver function tests, thyroid function tests, and other evaluations may be carried out. Chest and skull X rays are made, and computerized axial tomography (CAT scan) is used to reveal focal or generalized brain disease. Electroencephalography may show localized abnormalities in the electrical conduction of the brain caused by a lesion. Detailed psychological testing may reveal more specific perceptual, memory, or other disabilities.

Senile and presenile dementia.

In these dementias there is a progressive intellectual impairment that proceeds to lethargy, inactivity, and gross physical deterioration and eventually to death within a few years. Presenile dementias are arbitrarily defined as those that begin in

persons under the age of 65. In old age the most common causes of dementia are Alzheimer’s disease and cerebral arteriosclerosis. Dementia from Alzheimer's disease usually begins in people over age 65 and is much more common in women than in men. It begins with incidences of forgetfulness, which become more frequent and serious, and the disturbances of memory, personality, and mood progress steadily toward physical deterioration and death within a few years. In dementia caused by cerebral arteriosclerosis there are multiple areas of destruction of the brain caused by pieces of the damaged lining of arteries outside the skull lodging in the small arteries of the brain. The course of the illness is stepwise, with rapid deterioration followed by periods of slight improvement. Death may be delayed slightly longer than with dementia from Alzheimer's disease and often occurs from ischemic heart disease (heart attack) or from massive cerebral infarction, causing a stroke.

Other causes of dementia include Pick's disease, a rare inherited condition that occurs in women twice as often as men, usually between the ages of 50 and 60; Huntington's chorea, an inherited disease that usually begins at about the age of 40 with involuntary movements and proceeds to dementia and death within 15 years; and Creutzfeldt-Jakob disease, a rare condition that is probably caused by a transmissible agent known as a slow virus. Head injury, for instance, resulting from a boxing career or from an accident, may produce dementia. Infection, for example, with neurosyphilis or encephalitis, various tumours, toxic conditions such as chronic alcoholism or heavy metal poisoning, metabolic illnesses such as liver failure, reduced oxygen to the brain due to anemia or carbon monoxide poisoning, and the inadequate intake or metabolism of certain vitamins may all result in dementia.

There is no specific treatment for the symptoms of dementia; the underlying physical cause needs to be identified and treated when possible. The aims in the care of the demented patient are to relieve distress, prevent behaviour that might result in accident, and optimize his remaining physical and psychological faculties.

Substance-induced organic disorders.

A variety of psychiatric conditions can result from the use of alcohol or other drugs. Mental disorders resulting from the ingestion of alcohol include intoxication, withdrawal, hallucinations, and amnesia. Similar syndromes may occur following the use of other drugs. Those most commonly used nonmedically to alter mood are barbiturates, opioids (such as heroin), cocaine, amphetamines, hallucinogens such as lysergic acid diethylamide (LSD), cannabis, tobacco, and caffeine. Treatment is directed at alleviating symptoms and preventing the patient's further abuse of the substance.

Other organic syndromes.

Delirium occurs secondarily to many other physical conditions such as drug intoxication or withdrawal, metabolic disorders (for example, liver failure or low blood sugar), infections such as pneumonia or meningitis, head injuries, brain tumours, epilepsy, or nutritional or vitamin deficiency. There are a clouding or confusion of consciousness and disturbances of thinking, behaviour, perception, and mood, with disorientation being prominent. Treatment is aimed at the underlying physical condition.

Damage to different areas of the brain may cause particular psychological symptoms. Damage to the frontal lobe of the brain may manifest itself in such disturbances of behaviour as loss of inhibitions, tactlessness, and overtalkativeness. Lesions of the parietal lobe may result in difficulties of speech and language or of the perception of space. Lesions of the temporal lobe may lead to emotional instability, aggressive behaviour, or problems with learning new information.

Substance use disorders.

This category refers to maladaptive behaviour associated with the regular nonmedical use of substances that affect the central nervous system. Substance abuse implies a sustained pattern of pathological use resulting in impairment of the drug abuser's social or occupational functioning. Substance dependence implies

tolerance, in which markedly increased amounts of the drug must be administered to achieve the same effect, and withdrawal, in which symptoms follow the cessation of drug use or decreases in the dose of the substance.

Schizophrenia.

The term schizophrenia was introduced by the Swiss psychiatrist Eugen Bleuler in 1911 to describe what he considered to be a group of severe mental illnesses with related characteristics; "schizophrenia" eventually replaced the earlier term dementia praecox, which the German psychiatrist Emil Kraepelin had first used in 1899 to distinguish the disease from manic-depressive psychosis. Schizophrenic patients have a wide variety of symptoms; thus, although different authorities may agree as to whether a particular patient suffers from the condition, they might disagree about which constellations of symptoms are essential in clinically defining schizophrenia.

In 12 very different countries, rates for schizophrenia have been found to be surprisingly similar, the annual prevalence, that is, the number of cases both old and new recorded in one year, being between two and four per 1,000 persons. The lifetime risk of developing the illness is between seven and nine per 1,000. Schizophrenia is the single largest cause of admissions to mental hospitals, and it accounts for an even larger proportion of the permanent populations of such institutions. It is a severe and frequently chronic illness that typically first manifests itself during the teen years or during early adult life. More severe levels of impairment and personality disorganization are reached in schizophrenia than in almost any other mental disorder.

Clinical features.

The principal clinical signs of schizophrenia are delusions, hallucinations, a loosening or incoherence of a persons's thought processes and train of associations, deficiencies in feeling appropriate or normal emotions, and a withdrawal from reality. A delusion is a false or irrational belief that is firmly held despite obvious

or objective evidence to the contrary. The delusions of schizophrenics may be persecutory, grandiose, religious, sexual, or hypochondriacal in nature, or they may be concerned with other topics. Delusions of reference, in which the patient attributes a special, irrational, and usually negative significance to people, objects, or events in relation to himself, are common in the disease. Especially characteristic of schizophrenia are delusions in which the patient believes his thinking processes, parts of his body, or his actions or impulses are controlled or dictated by some external force. Schizophrenic delusions are frequently bizarre or absurd.

Hallucinations are false sensory perceptions that are experienced without an external stimulus but that nevertheless seem real to the subject. Auditory hallucinations, experienced as "voices" and characteristically heard commenting negatively about the patient in the third person, are prominent in schizophrenia. Hallucinations of touch, taste, smell, and bodily sensation also occur. Disorders of thinking vary in nature but are quite common in schizophrenia. The thought disorders may consist of a loosening of associations, so that the speaker jumps from one idea or topic to another unrelated one in an illogical, inappropriate, or disorganized way. At its most serious, this incoherence of thought extends into pronunciation itself, and the speaker's words become garbled or unrecognizable. Speech may also be overly concrete and inexpressive; it may be repetitive, or, though voluble, it may convey little or no real information. Usually a schizophrenic patient has little or no insight into his own condition and realizes neither that he is suffering from mental illness nor that his thinking is disordered.

Among the so-called negative symptoms of schizophrenia are a blunting or flattening of the person's ability to experience (or at least to express) emotion, indicated by speaking in a monotone and by a peculiar lack of facial expressions. The person's sense of self (i.e., of who he is) may be disturbed. He may be apathetic and may lack the drive and ability to pursue a course of action to its logical conclusion, or he may withdraw from the world, become detached from

others, and become preoccupied with silly, bizarre, or nonsensical fantasies. Such symptoms are more typical of chronic rather than of acute schizophrenics.

Different authorities have recognized many different types of schizophrenia, and there are intermediate stages between the disease and other conditions. Four major types of schizophrenia are still recognized by the DSM-III: the disorganized or hebephrenic type, the catatonic type, the paranoid type, and the simple or undifferentiated type. Hebephrenic schizophrenia is characterized by grossly inappropriate, shallow, or silly emotional responses and by incoherent thought and speech. Catatonic schizophrenia is marked by striking motor behaviour, such as remaining motionless in a rigid posture for hours or days, and by stupor or mutism. Paranoid schizophrenia is marked by the presence of prominent delusions of a persecutory and/or grandiose nature. Undifferentiated schizophrenia is marked by an insidious or gradual reduction in the person's interest in and relations with the external world and by a pervasive impoverishment of his personality and emotional responses.

Course and prognosis.

The course of schizophrenic illness is extremely variable. It may be said that roughly one-third of schizophrenic patients make a complete recovery and have no further recurrence, one-third have recurrent episodes of the illness, and one-third deteriorate into chronic schizophrenia with severe disability. The prognosis for schizophrenics has improved during the 20th century due to the use of antipsychotic drugs and community supportive measures.

About 10 percent of schizophrenic patients die by suicide. The prognosis of schizophrenia is poor when it has a gradual rather than a sudden onset, when the patient is quite young at onset, when there is a long duration of illness, when the patient exhibits blunted feelings or has displayed an abnormal personality previous to the onset of the disease, and when such social factors as never having been married, poor sexual adjustment, a poor work record, or social isolation exist in the patient's personal history.

Etiology.

An enormous amount of research has been carried out to try to determine the causes of schizophrenia. Family, twin, and adoption studies provide strong evidence to support an important genetic contribution, but the mode of inheritance is not known. Stressful life events are known to trigger or quicken the onset of schizophrenia or to cause relapse. Some abnormal neurological signs have been found in schizophrenics, and it is possible that brain damage, perhaps occurring at birth, may be a cause in some cases. Various biochemical abnormalities have been reported in schizophrenics, but the evidence for the causal relevance of these abnormalities is incomplete.

Much research has been carried out to determine whether the types of communication used in the families of schizophrenics or the parental care in such families help produce the disease. There has also been extensive interest in such factors as social class, place of residence, migration, and social isolation. Neither family dynamics nor social disadvantage have been proved to be causative agents.

Treatment.

The most successful treatment approaches combine the use of drugs, psychotherapy, and supportive therapy. In acute schizophrenia, phenothiazine, chlorpromazine, or butyrophenone drugs such as haloperidol are of proven efficacy in relieving or eliminating such symptoms as delusions, hallucinations, thought disorders, agitation, and violent behaviour. Long-term maintenance on such drugs also reduces the rate of relapse. Psychotherapy serves to relieve the patient's feelings of helplessness and isolation, buttress his healthy or positive tendencies, and help him to distinguish between his psychotic perceptions and reality and to deal with any underlying emotional conflicts that might be exacerbating his condition. Occupational therapy for those in day care and regular visits from a social worker or community psychiatric nurse for outpatients are beneficial. It is sometimes useful to counsel the relatives of schizophrenic patients living at home in their way of dealing with the patient's symptoms.

Paranoid disorders.

Paranoia is a syndrome in which a person thinks or believes, without justification, that other people are plotting or conspiring against him, are harassing him, or are otherwise persecuting or trying to harm him in some way. Paranoid thinking frequently causes a person to interpret or exaggerate innocuous or trivial incidents in a self-referrent way; e.g., to see two people talking at a distance and to irrationally assume that they are plotting against or criticizing him. Grandiosity or delusions of grandeur, which consist of exaggerated and unjustified ideas of a person's own importance, wealth, or power, frequently coexist with the classic persecutory orientation in paranoia. Paranoia or paranoid thinking can be a prominent or primary feature in schizophrenia (paranoid schizophrenia), personality disorders, senile dementias, affective disorders, and manic-depressive psychoses, and indeed it is difficult to demarcate strictly what the DSM-III defines as paranoid disorders proper. Persons with paranoid disorders may be otherwise normal people who are simply abnormally suspicious, or they may have an unshakable and highly elaborate delusional system involving worldwide conspiracies against them. A special type of paranoia is delusional jealousy, in which a person delusionally believes or suspects that his spouse is having sexual relations with someone else. A paranoid disorder can seriously impair an individual's social or marital functioning, but his thinking remains clear and orderly, his intellectual functioning is impaired only minimally or not at all, and the core of his personality remains intact. Many people with paranoid disorders can have normal or near-normal careers. The treatment of persons with paranoid disorders involves the use of antipsychotic drugs, frequently on a long-term maintenance basis.

Affective disorders.

These disorders are usually restricted to just two abnormalities of mood - depression and elation, or mania. (Mood is a predominant emotion that colours the individual's entire psychic life.)

Depression is characterized by a sad or hopeless mood, pessimistic thinking, a loss of enjoyment and interest in one's usual activities and pastimes, reduced energy and vitality, increased fatigue, slowness of thought and action, loss of appetite, and disturbed sleep or insomnia. Depression must be distinguished from the grief and low spirits felt in reaction to the death of a loved one or some other unfortunate circumstance. The most dangerous consequence of severe depression is suicide.

Mania is characterized by an elated or euphoric mood, quickened thought and accelerated, loud or voluble speech, overoptimism and heightened enthusiasm and confidence, inflated self-esteem, heightened motor activity, irritability, excitement, and a decreased need for sleep. The manic individual may become injured, commit illegal acts, or suffer financial losses due to the poor judgment and risk-taking behaviour he displays when in the manic state.

There are enormous problems in the classification of affective disorders, particularly of depression, and the various clinical distinctions made by different authorities are difficult to correlate with particular sets of symptoms or particular causes. An important distinction, however, is made between depressions that are endogenous (i.e., arising independently of environmental influences and presumably caused by a biochemical imbalance) and those that are reactive (i.e., arising in response to external stresses or trauma).

Major affective disorders.

The DSM-III defines two major, or severe, affective disorders: bipolar disorder and major depression. A person with bipolar disorder, which has traditionally been called manic-depressive psychosis, typically experiences discrete episodes of depression and then of mania lasting for a few weeks or months, with intervening periods of complete normality. The sequence of depression and mania can vary extremely from patient to patient and within one individual, with either mood abnormality predominating in duration and intensity. Depressive mood swings typically occur more often and last longer than manic ones, though there are patients who have episodes only of mania. Patients with bipolar disorder frequently

also show such psychotic symptoms as delusions, hallucinations, paranoia, or grossly bizarre behaviour.

The lifetime risk for developing bipolar disorder is about 0.7 percent and is about the same for men and women. The onset of the illness often occurs around the age of 30, and the illness persists over the long term. The predisposition to develop bipolar disorder is partly genetically inherited. Antipsychotic drugs such as chlorpromazine or haloperidol are used for the treatment of acute mania. Lithium carbonate has proved effective in both treating and preventing recurrent attacks of mania.

Severe and long-lasting depression without the presence of mania is classified by the DSM-III as major depression. Depression is a much more common illness than mania, and there are indeed many sufferers from depression who have never experienced mania. Major depression may occur as a single episode or it may be recurrent. It may also exist with or without melancholia and with or without psychotic features. Melancholia implies the so-called biological symptoms of depression: early-morning waking; daily variations of mood with depression most severe in the morning; loss of appetite and weight; constipation; and loss of interest in love and sex. Melancholia is a particular depressive syndrome that is relatively more responsive to physical methods of treatment, such as drugs and electroconvulsive therapy.

It is estimated that the annual incidence of major depression is about 140 for men and 4,000 for women per 100,000 population. While the rates for major depression in men increase with age, the peak for women is between the ages of 35 and 45. There is a serious risk of suicide with the illness; of those who have a severe depressive disorder, about one-sixth eventually kill themselves. The loss of one's parents or other childhood traumas or deprivations can increase a person's vulnerability to depression later in life, and stressful life events, especially where some type of loss is involved, are in general potent precipitating causes of the illness. It seems that both psychosocial and biochemical mechanisms are important in causing depression. Of the latter factor, the best-supported hypotheses suggest

that the faulty regulation of the release of one or more naturally occurring amines at sites in the brain where the transmission of nerve impulses takes place is the basic cause, with a deficiency of the amines resulting in depression and an excess causing mania. The most likely candidates for the suspect amines are the biological monoamines (norepinephrine, dopamine, and 5-hydroxytryptamine). The treatment of major depressive episodes usually requires antidepressant drugs; electroconvulsive therapy may also be helpful, as may cognitive psychotherapy.

Minor affective disorders.

A less severe manifestation of the manic-depressive syndrome, in which the mood swings are present but not as extreme, is termed cyclothymic disorder. This illness is better considered a personality disorder of affective type; the prevailing mood swings are established in adolescence and continue throughout adult life.

Dysthmic disorder, or depressive neurosis, may occur on its own, but it more commonly appears along with other neurotic symptoms such as anxiety, phobia, and hypochondriasis. Where there are clear external grounds for a person's unhappiness, a dysthymic disorder is considered to be present when the depressed mood is disproportionately severe or prolonged in regard to the distressing experience, when there is a preoccupation with the precipitating situation, when the depression continues even after removal of the provocation, and when it impairs the individual's ability to cope with the specific stress.

At any time, depressive symptoms may be found in one-sixth of the population, more commonly in women than men. Social factors are important etiologically, as evidenced in the high rates of depression found in urban women living without a male cohabitant, having three or more children, and lacking employment outside the home. Loss of self-esteem, feelings of helplessness and hopelessness, and losses of various types of "loved objects" are also seen as important causes of minor depression. The course and severity of dysthymic disorder is extremely variable - from a few weeks or months to several decades and from the mild

impairment of social functioning to almost total incapacitation. Psychotherapy is the treatment of choice, although antidepressant medication may prove beneficial.

Anxiety disorders.

Anxiety has been defined as a feeling of fear, dread, or apprehension that arises without a clear or appropriate real-life justification. Some authorities differentiate anxiety from true fear in that the latter is experienced in response to an actual threat or danger, such as those to one's physical safety. Anxiety, on the other hand, may arise in response to apparently innocuous situations or may be out of proportion to the actual degree of the external stress. Anxiety also frequently arises as a result of subjective emotional conflicts of whose nature the person himself may be unaware. Generally, intense, persistent, or chronic anxiety that is not justified in response to real-life stresses and that interferes with the individual's functioning is regarded as a manifestation of mental disorder. Anxiety is a symptom in many mental disorders, including schizophrenia, obsessive-compulsive disorders, posttraumatic stress disorders, and so on, but in phobias and other anxiety disorders proper, anxiety is the primary and frequently the only symptom.

The symptoms of anxiety are physical, psychological, and behavioral. Anxiety, especially during panic attacks, can manifest itself in a distinctive set of physical signs that arise from overactivity of the sympathetic nervous system or from tension in skeletal muscles. The sufferer experiences palpitations, dry mouth, dilatation of the pupils, shortness of breath, sweating, abdominal symptoms, tightness in the throat, trembling, and dizziness. Aside from the actual feelings of dread and apprehension, the psychological symptoms include irritability, difficulty with concentration, and restlessness. Anxiety may also be manifested in avoidance behaviour - running away from the feared object or situation.

Phobic disorder or neurosis.

Phobias are neurotic states accompanied by intense dread of certain objects or situations that would not normally have such an effect. This type of anxiety is associated with a strong desire to avoid the dreaded object or situation. About six per 1,000 of the population suffer from a phobic disorder. There is a tendency for phobic symptoms, whatever their nature, to persist for many years unless treated, and the avoidance behaviour they produce can seriously limit the affected individual's movements and his social or occupational functioning.

People can have phobias about many different kinds of objects or situations, but three main divisions of phobic syndromes are made by the DSM-III: simple phobia, agoraphobia, and social phobia. Individuals with simple phobias may intensely fear a specific object or situation, for example, cats or thunderstorms; they have anxious thoughts upon anticipating contact with an object or event, for instance, upon hearing the weather forecast, and they try to avoid the object, as in staying indoors in order not to encounter a cat. Typically, agoraphobic patients have an intense fear of being alone in or being unable to escape from a public place or some other setting outside the home, such as a crowded bus or a supermarket. A social phobia is present when the individual has extreme anxiety in a social situation where he is under the scrutiny of others, such as eating in a restaurant or speaking at a meeting.

The treatment of phobic disorders is best approached by the use of behavioral therapy; dynamic psychotherapy and antianxiety drugs may be effective in some cases.

Anxiety states or neuroses.

Anxiety disorders in which the anxiety is not aroused by any specific object or situation can basically be subsumed under the headings of panic disorder and generalized anxiety disorder. Panic attacks are characterized by the sudden onset of intense or overwhelming anxiety accompanied by any of the aforementioned physical signs, such as difficulty in breathing, sweating, palpitations, and so on. The fear and apprehension experienced in such attacks sometimes mount to what

are known as feelings of doom. Clear precipitating circumstances may produce the initial feelings of intense anxiety. The panic attack may last for about a quarter of an hour and frequently recurs, either infrequently or several times a week. The disorder usually starts in young adults and may persist for many years.

A diffuse and persistent feeling of anxiety associated with no particular object or situation is termed general, or free-floating, anxiety and is classified by the DSM-III as generalized anxiety disorder. General anxiety is usually milder and less intense than in panic attacks, but it is longer lasting and may persist for several months or years, or on a recurrent basis. The most effective treatments vary according to the type of disorder and the individual patient. Psychotherapy and antianxiety drugs are often useful in treating generalized anxiety and panic attacks.

Obsessive-compulsive disorder or neurosis.

In this condition an individual experiences obsessions or compulsions or both. Obsessions are recurring words, thoughts, ideas, or images that, rather than being experienced as voluntarily produced, seem to invade a person's consciousness despite his attempts to ignore, control, or suppress them. The obsessional thought or topic is perceived by the sufferer as inappropriate or senseless; the idea is recognized both as alien to his nature and yet as coming from inside himself. An obsession can take the form of a recurrent and vivid fantasy that is often obscene, disgusting, repugnant, or senseless. The patient with obsessional ruminations holds endless debates over mundane matters inside his head; e.g., "Did I forget to lock the front door behind me?"

Obsessions in turn are frequently linked to compulsions. These are urges or impulses to perform repetitive acts that are apparently meaningless, unnecessary, stereotyped, or ritualistic. The compulsive person knows that the act to be performed is meaningless or unnecessary, but his failure or refusal to perform it brings on a mounting tension or anxiety that is temporarily relieved once the act is performed. Obsessional ruminations thus directly produce compulsive behaviour; e.g., repeatedly checking and relocking an already secure front door. Most

compulsive acts have a simple, ritualistic character and can involve checking, touching, hand-washing, or the repetition of particular words or phrases.

Drugs, psychotherapy, and behavioral therapy are selectively successful in treating obsessive-compulsive disorders, depending on the individual patient. The drug clomipramine has proved to be notably effective in reducing or even eliminating the symptoms in a large proportion of patients tested.

Posttraumatic stress disorder.

In this condition symptoms develop in an individual after he has experienced a psychologically traumatic event. It is a category in the DSM-III classification but is not different in its symptomatology from certain other neurotic conditions; the distinctive feature is the presence of external trauma. The traumatic events can include serious automobile accidents, rape or assault, military combat, torture, incarceration in a concentration or death camp, and such natural disasters as floods, fires, or earthquakes.

A feature of this condition is the person's reexperiencing of the traumatic event in nightmares and in intrusive daytime fantasies. Sometimes an insignificant event, like a knock at the door, will precipitate a sudden terrifying recollection and an exaggerated startle response. Other symptoms include emotional numbing, a diminished ability to enjoy activities or relationships that were previously pleasurable, and difficulty with sleeping. Long-term symptoms of distress, marital and family problems, difficulties at work, and the abuse of alcohol and other drugs are characteristic impairments caused by the disorder.

The marked emotional symptoms may persist long after the traumatic event actually occurs. Some persons are more liable than others to develop the disorder, depending on personality traits, previous psychological disturbances, age, and genetic predisposition. Psychotherapy is the basic approach used in treating this disorder.

Somatoform disorders.

In these conditions, the physical symptoms of the person suggest the presence of organic disease but no such organic disorder can be found upon physical examination and investigation, and instead there is positive evidence that the symptoms are caused by psychological factors. The production of these symptoms is not under voluntary control. The terms hypochondriasis and hysteria that traditionally designated these disorders are still widely used by psychiatrists.

Somatization disorder.

This disorder was previously designated Briquet's syndrome; its essential features consist of multiple, recurrent physical complaints made over many years and starting in young adult life or adolescence. The sufferer demands medical attention, but no organic cause is found. The symptoms invariably occur in many different bodily systems, for instance back pains, painful menstruation, dizziness, indigestion, difficulty with vision, and partial paralysis; and the symptoms may follow fashions in health concerns among the public.

The condition is relatively common and occurs in about 1 percent of adult women. It is very unusual to see this disorder in males. There are no clear etiological factors. Treatment involves not colluding with the patient's inclination to attribute organic causes to the symptoms and insuring that physicians and surgeons do not cooperate with the patient in seeking excessive diagnostic procedures or surgical remedies for the complaints.

Conversion disorder or hysterical neurosis, conversion type.

This disorder was traditionally labeled hysteria. Its symptoms are a loss of or alteration in physical functioning, typically the paralysis suggesting neurological disease. The physical symptoms occur in the absence of organic pathology and are instead apparently the expression of an underlying emotional conflict. The characteristic motor symptoms of hysteria include the paralysis of the voluntary muscles of an arm or leg, tremor, tics, and other disorders of movement or gait. The neurological symptoms may be widely distributed and may not conform with

medical knowledge of physical nerve distribution. Blindness, deafness, loss of sensation in arms or legs, the feeling of "pins and needles," an increased sensitivity to pain in a limb, and many other symptoms have been described.

Hysterical symptoms usually occur in a setting of extreme psychological stress and appear suddenly. The course is variable, with recovery often occurring in a few days but with symptoms persisting for years or decades in chronic cases that remain untreated.

The causation of hysteria has been linked with fixations; i.e., arrested stages in the individual's early psychosexual development. Freud's theory that threatening or emotionally charged thoughts are repressed out of consciousness and converted into physical symptoms is still widely accepted. The treatment of hysteria thus requires psychological rather than pharmacological methods, notably the exploration of the sufferer's underlying emotional conflicts. Hysteria (and hypochondriasis) can also be considered as different forms of "illness behaviour"; i.e., the patient uses the hysterical symptoms to gain a psychological advantage in social relationships, either by gathering sympathy or by being relieved of burdensome or stressful obligations and withdrawing from emotionally disturbing or threatening situations. Thus it may be advantageous to the patient, in a psychological sense, to have the consequences of the symptoms.

Hypochondriasis or hypochondriacal neurosis.

Hypochondriasis is a preoccupation with physical signs or symptoms that the patient unrealistically interprets as abnormal, leading to the fear or belief that he is seriously ill. There may be fears about the development of physical or mental symptoms without any such existing, a belief that actual but minor symptoms are of dire consequence, or an experience of normal bodily sensations as threatening symptoms. A thorough physical examination may find no organic cause for the physical signs the patient is concerned about, but the examination fails to relieve his unrealistic fears about having a serious disease. The symptoms of

hypochondriasis may occur with mental illnesses other than neuroses, for instance, depression or schizophrenia.

Hypochondriacal neurosis occurs in both sexes. The onset may be associated with precipitating factors such as an actual organic disease with physical and psychological aftereffects; e.g., coronary thrombosis in a previously fit man. It often begins during the fourth and fifth decades of life but is also common at other times, during pregnancy, for example. Treatment aims to provide understanding and support and to reinforce healthy behaviour; antidepressant drugs may be used when there are depressive symptoms.

Other somatoform disorders.

In psychogenic pain disorder the main feature is the persistent complaint of pain in the absence of organic disease and with evidence of a psychological cause. The pattern of pain may not conform to the known anatomic distribution of the nervous system. Psychogenic pain may occur as part of hypochondriasis or as a symptom of a depressive disorder. Appropriate treatment depends on the context of the symptom.

These somatoform disorders may occur together in one patient. Alternatively, they may occur in atypical form or in association with another physical or mental illness.

Dissociative disorders or hysterical neurosis, dissociative type.

Dissociation is a syndrome in which one or a group of mental processes are split off, or dissociated, from the rest of the psychic apparatus so that their function is lost, altered, or impaired. Dissociative symptoms have often been regarded as the mental counterparts of the physical symptoms displayed in conversion disorders. Since the dissociation may be an unconscious mental attempt to protect the individual from threatening impulses or emotions that are repressed, the conversion into physical symptoms and the dissociation of mental processes can be seen as related defense mechanisms arising in response to emotional conflict.

In dissociative disorders there is a sudden, temporary alteration in the person's consciousness, sense of identity, or motor behaviour. There may be an apparent loss of memory of previous activities or important personal events, with amnesia for the episode itself after recovery. These are rare conditions, and it is important to exclude organic causes.

In psychogenic or hysterical amnesia there is a sudden loss of memory which may appear total; the patient can remember nothing about his previous life or even his name. The amnesia may be localized to a short period of time associated with a traumatic event or it may be selective, affecting the person's recall of some, but not all, of the events during a particular time. In psychogenic fugue, the individual wanders away from his home or place of work and assumes a new identity; he cannot remember his previous identity and upon recovering cannot recall the events that occurred while he was in the fugue state. In many cases the disturbance lasts only a few hours or days and involves only limited travel. Severe stress frequently triggers this disorder.

Multiple personality is a rare and remarkable dissociative disorder in which two or more distinct and independent personalities develop in a single individual. Each of these personalities inhabits the person's conscious awareness to the exclusion of the others at particular times. This disorder frequently arises as a result of traumas suffered during childhood and is best treated by psychotherapy, which seeks to reunite the various personalities into a single, integrated one.

In depersonalization a person feels or perceives his body or self as being unreal, strange, altered in quality, or distant. This state of self-estrangement may take the form of feeling as if one is machinelike, is living in a dream, or is not in control of one's actions. Derealization, or feelings of unreality concerning objects outside one's self, often occurs at the same time. Depersonalization may occur alone in neurotic patients but is more often associated with phobic, anxiety, or depressive symptoms. It most commonly occurs in younger married women and may persist for many years. Patients find the experience of depersonalization intensely difficult to describe and often fear that people will think them insane. Organic conditions,

especially temporal lobe epilepsy, must be excluded before making a diagnosis of neurosis when depersonalization occurs. As with other neurotic syndromes, it is more common to see a mixed picture with many different symptoms than depersonalization alone.

The causes of depersonalization are obscure, and there is no specific treatment for it. When the symptom arises in the context of another psychiatric condition, treatment is aimed at that illness.

Personality disorders.

Personality is the characteristic way in which an individual thinks, feels, and behaves; it accounts for the ingrained behaviour patterns of the individual and allows the prediction of how he will act in particular circumstances. Personality embraces a person's moods, attitudes, and opinions, and is most clearly expressed in his interactions with other people. A personality disorder is a deeply ingrained, long-enduring, maladaptive, and inflexible pattern of thinking, feeling, and behaving that either significantly impairs an individual's social or occupational functioning or causes him subjective distress. Personality disorders are not illnesses but rather are pronounced accentuations or variations of personality in one or more of its traits.

A personality disorder may occur with another psychiatric condition or on its own, and it is particularly likely to be associated with neurotic conditions. The causes of personality disorders are obscure. There is undoubtedly a constitutional and therefore hereditary element in determining personality type. Psychological and environmental factors are also important in causation, for instance, the association of antisocial personality disorders with other features of social deviance found in some families and in members of lower socioeconomic groups.

Some generally accepted types of personality disorder are listed below. It is important to recognize that simply exhibiting the trait or even having it to an abnormal extent is not enough to constitute disorder - for that, the degree of abnormality must cause disturbance to the individual or to society.

Paranoid personality disorder.

In this disorder there is a pervasive and unjustified suspiciousness and mistrust of others, whose words and actions are misinterpreted as having special significance for, and as being directed against, the individual. Sometimes such people are guarded, secretive, aggressive, quarrelsome and litigious, and excessively sensitive to the implied criticism of others.

Affective personality disorder.

Three particular types of persistent mood disturbance can be described under this heading: (1) the trait of anxiety may be persistent and highly developed, so that the person encounters all new circumstances with fearful anticipation; (2) the chronic depressive personality is a gloomy pessimist who is skeptical in outlook and who may regard suffering as meritorious; and (3) the cyclothymic personality shows excessive swings of mood as a persistent lifelong trait.

Schizoid personality disorder.

In this disorder there is a disinclination to mix with others, the individual appearing aloof, withdrawn, indifferent, unresponsive, and disinterested. Such a person prefers solitary to gregarious pursuits, involvement with things rather than with people, and often appears humourless or dull.

Schizotypal personality disorder.

This category has been used to describe people who show various oddities or eccentricities of thought, speech, perception, or behaviour (such as bizarre fantasies or persecutory delusions) but whose symptoms are not severe enough to be labeled as schizophrenic.

Explosive personality disorder.

Such people have a tendency to sudden emotional rages or tantrums that result in their physically assaulting others or impulsively attempting to commit suicide. The

emotional outburst may be precipitated by a minor frustration that is disproportionate to the degree of reaction.

Anankastic, or compulsive, personality disorder.

A person with this disorder shows prominent overscrupulous, perfectionistic traits that are expressed in feelings of insecurity, self-doubt, meticulous conscientiousness, indecisiveness, and rigidity of behaviour.

The person is preoccupied with rules, procedures, and efficiency, is overly devoted to work and productivity, and is usually deficient in the ability to express warm or tender emotions. This disorder is more common in men and is in many ways the antithesis of antisocial personality disorder.

Histrionic personality disorder.

Overly dramatic and intensely expressed behaviour, a tendency to call attention to oneself, a craving for novelty and excitement, egocentricity, highly reactive and excitable behaviour, and tendencies toward dependency and suggestibility are characteristic of this condition, which is more common in women than men.

Asthenic, dependent, or inadequate personality disorder.

In this condition the person lacks the mental energy and ability to act on his own initiative and therefore passively allows others to assume responsibility for major aspects of his life.

Antisocial, asocial, sociopathic, or psychopathic personality disorder.

This disorder is marked by a personal history of chronic and continuous antisocial behaviour, in which the rights of others are violated, and by poor or nonexistent job performance. It is manifested in persistent criminality, sexual promiscuity or aggressive sexual behaviour, and drug use. People with this disorder are impulsive, mendacious, irresponsible, and callous; they feel no guilt over their antisocial acts and fail to learn from their mistakes. The symptoms usually appear in adolescence.

Antisocial personalities are less liable to criminal acts as they grow older, but there remains a high risk of suicide, accidental death, drug or alcohol abuse, and a tendency toward interpersonal problems.

Other categories of personality disorder.

In the narcissistic personality disorder, there is a grandiose sense of self-importance and a preoccupation with fantasies of success, power, and achievement. Avoidant personalities are excessively sensitive to social rejection, humiliation, and shame, have low self-esteem, and are deeply upset by the slightest disapproval of others; they are consequently unwilling to enter into relationships but crave affection and acceptance. Passive-aggressive personality disorder is the term applied to people who respond aggressively and negatively to demands made upon them by using such passive means as procrastination, dawdling, intentional inefficiency, or deliberate forgetfulness.

Personality traits are, by definition, virtually permanent, and so these disorders are only partially, if at all, amenable to treatment. The most effective treatment combines various types of group, behavioral, and cognitive psychotherapy. The behavioral manifestations of personality disorders often tend to diminish in their intensity in middle and old age.

Psychosexual disorders.

Transsexualism and disorders of gender identity.

In transsexualism the person feels a discrepancy between anatomical sex and the gender the person ascribes to himself. This disorder is much more common in biological males than females. The sufferer claims that he is a member of the other sex: "a female spirit trapped in a male body." He may assume the dress and behaviour and participate in activities commonly associated with the other sex and may even use hormones and surgery to achieve "restitution to my rightful appearance"; i.e., to achieve the physical characteristics of the other sex. The cause of the condition is unknown. Once established, transsexualism persists for many

years, perhaps for the rest of life. There is a risk of developing depression and an increased risk of suicide. Psychiatric treatment is generally supportive in type.

Paraphilias.

Paraphilias, or sexual deviations, may be classified into disorders of sexual object and of the sexual act. Disorders of sexual object include the following. (1) In fetishism, inanimate objects are the repeated sexual preference and means of sexual arousal. (2) In transvestism, the recurrent wearing of clothes of the opposite sex is carried out to achieve sexual excitement. (3) In zoophilia, or bestiality, an animal is used as the repeated and preferred means of achieving sexual excitement. (4) In pedophilia, an adult has sexual fantasies about or engages in sexual acts with a prepubertal child of the same or opposite sex.

Disorders of the sexual act include the following. (1) In exhibitionism, repeated exposure of the genitals to an unsuspecting stranger is used to achieve sexual excitement. (2) In voyeurism, observing the sexual activity of others repeatedly is the preferred means of sexual arousal. (3) In sexual masochism, the individual achieves sexual excitement from being made to suffer. (4) In sexual sadism, the individual achieves sexual excitement by inflicting suffering upon another person.

There are, of course, other unusual sexual objects or acts that may be used for gratification. The causes of these conditions are generally not known. Behavioral, psychodynamic, and pharmacological methods have been used with varying efficacy to treat these disorders.

Disorders usually first evident in infancy, childhood, or adolescence.

Children are usually referred to a psychiatrist or therapist because of complaints or concern over the child's behaviour or development by a parent or some other adult. Family problems, particularly difficulties in the parent-child relationship, are often an important causative factor in the symptomatic behaviour of the child. For the practice of child psychiatry, the observation of behaviour is especially important as the child may not be able to express his feelings in words. Isolated psychological

symptoms are extremely common in children, but in one survey, disturbance amounting to psychiatric disorder was found to be present in 7 percent of all 10 and 11 year olds; boys were affected to twice the extent of girls.

Attention disorders.

Children with these disorders show a degree of inattention and impulsiveness that is markedly inappropriate for their stage of development. Gross overactivity in children has many causes, including anxiety, conduct disorder, or the effects of living in institutions. One type of overactivity, the hyperkinetic, or hyperactive, syndrome, is characterized by extreme restlessness and by sustained and prolonged motor overactivity such as running around. Learning difficulties and antisocial behaviour may occur secondarily. This syndrome is 10 times more common in boys than in girls.

Conduct disorders.

These are the most common psychiatric disorders in older children and adolescents, accounting for nearly two-thirds of disorders in those aged 10 and 11 years. Abnormal conduct more serious than ordinary childlike mischief persistently occurs; lying, disobedience, and aggression may be shown at home, and truancy, delinquency, and deterioration of work may occur at school. Vandalism, drug and alcohol abuse, and early sexual promiscuity may also occur. The most important causative factors are the family background; broken homes, unstable and rejecting families, institutional care in childhood, and a poor social environment are frequently present in such cases.

Anxiety disorders.

Neurotic or emotional disorders in children are similar to the adult conditions except that they are often less clearly differentiated. In anxiety disorders of childhood, the child is fearful, timid with other children, and overdependent and clinging toward the parents. Aches and other physical symptoms, sleep

disturbance, and nightmares occur. Separation from the parent or from the home environment is a major cause of this neurotic anxiety.

Eating disorders.

Anorexia nervosa usually starts in late adolescence and is about 20 times more common in girls than boys. This disorder is characterized by a body weight more than 25 percent below standard,

amenorrhea, a fear of loss of control of eating, and an intense desire to be thin. Though grossly thin, patients nevertheless believe themselves to be fat. They go to enormous lengths to resist eating food and to lose weight, including food avoidance, purging, self-induced vomiting, and vigorous exercise.

The condition appears to start with the patient's voluntary control of food intake in response to social pressures such as peer conformity. The disorder is exacerbated by troubled relations within the family. It is much more common in developed, wealthy societies and in girls of higher socioeconomic class. There is evidence that it has become more common in such countries since the 1960s. Patient management includes three stages: persuading the patient to accept and cooperate with treatment, achieving weight gain by medical methods of care, and helping the patient maintain weight by psychological and social therapy. Bulimia nervosa refers to episodic grossly excessive overeating binges. These may alternate with episodes of self-induced vomiting. The disorder is a variant of anorexia nervosa.

Disorders with physical manifestation.

Stereotyped movement disorders involve the exhibition of tics in differing patterns. A tic is an involuntary, purposeless jerking movement of a group of muscles or the involuntary production of noises or words. Tics may affect the face, head, and neck or, less commonly, the limbs or trunk. Gilles de la Tourette's syndrome is typified by multiple tics and involuntary vocalization, especially the uttering of obscenities.

Other physical symptoms that are often listed among psychiatric disorders of childhood include stuttering, enuresis (the repeated involuntary voiding of urine by

day or night), encopresis (the repeated voiding of feces into inappropriate places), sleepwalking, and night terror. These symptoms are not necessarily evidence of emotional disturbance or of some other mental illness.
Behavioral methods of treatment may sometimes be effective.

Infantile autism.
Psychotic disorders are very rare in childhood, and of these about one-half are cases of infantile autism; boys are affected three times as often as girls. Infantile autism begins in the first two years of life and is more common in the upper socioeconomic classes. The child shows an inability to make warm emotional relationships, has a severe speech and language disorder, and exhibits a desire for sameness in which he shows distress if thwarted from his stereotyped behaviour. There is some evidence to support genetic and organic factors in causation. Treatment involves management of the abnormal behaviour, training in life skills and occupational activities, and counseling for the family.

Other mental disorders.
Factitious disorders. These are characterized by physical or psychological symptoms that are voluntarily self-induced; they are distinguished from hysteria, in which the physical symptoms are produced unconsciously. In factitious disorders, although the person's attempts to create or exacerbate the symptoms of an illness are voluntary, such behaviour is neurotic in that the individual is unable to refrain from it; i.e., his goals, whatever they may be, are involuntarily adopted. In malingering, by contrast, the person simulates or exaggerates an illness or disability to obtain some kind of discernible personal gain or to avoid an unpleasant situation; e.g., a prison inmate may simulate madness to obtain more comfortable living conditions. It is important to recognize factitious disorders as evidence of psychological disturbance. Treatment is of the underlying conflicts.

Disorders of impulse control.

These conditions are usually associated with a disorder of personality. There is a failure to resist desires, impulses, or temptations to perform an act that is harmful to the individual or to others. The individual experiences a feeling of tension before committing the act and a feeling of release or gratification upon completing it. The behaviours involved include pathological gambling, setting fires (pyromania), and impulsive stealing (kleptomania).

Adjustment disorders.

These are neurotic conditions in which there is an inappropriate reaction to an external stress occurring within three months of the stress. The symptoms may be out of proportion to the degree of stress, or they may be maladaptive in the sense that they prevent the individual from coping adequately in his social or occupational setting. These disorders are often associated with other neurotic conditions such as anxiety neurosis or minor depression.

Treatment of mental disorders

HISTORICAL OVERVIEW

Early history.

References to mental disorders in early Egyptian, Indian, Greek, and Roman writings show that the physicians and philosophers who contemplated problems of human behaviour regarded mental illnesses as a reflection of the displeasure of the gods or as evidence of demoniac possession. Only a few realized that sufferers from mental illnesses should be treated humanely rather than exorcised, punished, or banished. Certain Greek medical writers, however, notably Hippocrates (flourished 400 BC), regarded mental disorders as diseases to be understood in terms of disturbed physiology. Hippocrates and his followers emphasized natural causes, clinical observation, and brain pathology in the study of mental disorders. Later Greek medical writers, including those who practiced in Imperial Rome, set out treatment programs for mental illness, including quiet, occupation, and the use

of drugs such as the purgative hellebore. It is probable that most psychotic people during ancient times were cared for by their families and that those who were thought to be dangerous to themselves or others were detained at home by relatives or hired keepers.

During the early Middle Ages in Europe, primitive thinking about mental illness reemerged, and witchcraft and demonology were used to account for the symptoms and behaviour of psychotic people. At least some of the insane were looked after by the religious orders, who offered care for the sick generally. The empirical and quasi-scientific Greek tradition in medicine was maintained not by the Europeans but by the Muslim Arabs, who are usually credited with the establishment of asylums for the mentally ill in the Middle East as early as the 8th century. In medieval Europe in general it seems that the madman was allowed his liberty, provided he was not regarded as dangerous. The founding of the first hospital in Europe devoted entirely to the care of the insane probably occurred in Valencia, Spain, in 1407-09, though this has also been said of a hospital established in Granada in 1366-67.

From the 17th century onward in Europe there was a growing tendency to isolate deviant people, including the insane, from the rest of society. Thus, such socially unwanted people as the mentally ill were confined together with the handicapped, vagrants, and delinquents. Those of the insane who were regarded as violent were often chained to the wall and were treated in a barbarous and inhumane way.

In the 17th and 18th centuries the development of European medicine and the rise of empirical methods of medicoscientific inquiry were paralleled by an improvement in public attitudes toward the mentally ill, which only began to emerge toward the end of that period. By the end of the 18th century, however, concern over the care of the insane had become so great among educated people in Europe and North America that governments were forced to act. After the French Revolution the physician Philippe Pinel was placed in charge of the Bicêtre, the hospital for the mentally ill in Paris. Under Pinel's supervision a completely new approach to the handling of mental patients was introduced. Chains and shackles

were removed from the patients, and in place of dungeons they were provided with sunny rooms and were permitted to exercise on the hospital grounds. Among other reformers were the British Quaker layman William Tuke, who established the York Retreat for the humane care of the mentally ill in 1796, and the physician Vincenzo Chiarugi, who published a humanitarian regime for his hospital in Florence in 1788. In the mid-19th century Dorothea Dix carried on a campaign to arouse the public to the inhumane conditions that prevailed in American mental hospitals, and her efforts led to widespread reforms both in the United States and elsewhere.

The mental hospital era.

Many hospitals for the insane were built in the latter half of the 18th century. Some of them, like the York Retreat in England, were run on humane and enlightened lines, while others, like the York Asylum, gave rise to great scandal because of their brutal methods and filthy living conditions. In the mid-19th century an extensive program of mental hospital building was carried out in North America, Britain, and many of the countries of continental Europe. The hospitals housed the insane poor, and their aim was to care for patients humanely and to relieve their families of the burden of caring for them. The approach was that of moral treatment, including occupation, the avoidance of physical methods of restraint, and respect for the individual patient. A widespread belief in the curability of mental illness at this time was a principal motivating factor behind such reform.

The mental hospital era was an age of reform, and there is no doubt that patients were treated much more humanely. The era produced a large number of segregated institutions in which a much higher proportion of the mentally ill were confined than previously. But the medical reformers' early hopes of successful cures were not vindicated, and by the end of the 19th century the hospitals had become overcrowded, and custodial care had replaced moral treatment.

The biological movement.

Along with humanitarian reforms in hospital practice and treatment methods during the late 18th and 19th centuries, there was a resurgence of medical and scientific interest in psychiatric theory and practice. Fundamental strides were made during this period in establishing a scientific basis for the study of mental disorders. A long series of observations by clinicians in France, Germany, and England culminated in 1883 in a comprehensive classification of mental disorders by the German psychiatrist Emil Kraepelin. His classification system served as the basis for all subsequent ones, and the cardinal distinction he made between schizophrenia and manic-depressive psychosis still stands.

Rapid advances in various branches of medicine led in the later 19th century to the expectation of discovering specific brain lesions that were thought to cause the various forms of mental disorder. While these researches did not attain the results that were expected, the scientific emphasis was productive in that it did elucidate the gross and microscopic pathology of many brain disorders that can produce psychiatric disabilities. Nevertheless, certain of the psychotic disorders, notably schizophrenia and manic-depressive psychosis, frustrated the effort to find causative agents in cellular pathology. It became apparent that other explanations had to be found for the many puzzling aspects of mental disorders in general, and these explanations emerged in a wave of psychological rather than physical explanations.

Development of psychotherapy.

Foremost among these was that of psychoanalysis, which originated in the work of the Viennese neurologist Sigmund Freud. Having studied under the French neurologist Jean-Martin Charcot, Freud originally used well-known techniques of hypnosis to treat patients suffering from hysterical paralysis and other neurotic syndromes. Freud and his colleague, Josef Breuer, observed that their patients tended to relive earlier life experiences that could be associated with the symptomatic expression of their illnesses. When these memories and the emotions associated with them were brought to consciousness during the hypnotic state, the

patients showed improvement. Observing that most of his patients proved able to talk about such memories without being under hypnosis, Freud evolved the technique of free association (the production by the patient, aloud and without suppression or self-censorship of any kind, of the thoughts and feelings about whatever was uppermost in his mind) as a means of access to the unconscious. From this beginning Freud gradually developed what became known as psychoanalysis. Other features of the new procedure included the study of dreams, the interpretation of "resistances" on the part of the patient, and the handling by the therapist of transference (the patient's feelings toward the analyst that reflected previously experienced feelings toward parents and other important figures in his early life). Freud's work, though complex and controversial in many of its aspects, laid the basis for modern psychotherapy in its use of free association and its emphasis on unconscious and irrational mental processes as causative factors in mental illness. This emphasis on purely psychological factors as a basis for both causation and treatment was to become the cornerstone of most subsequent psychotherapies.

Variations of the original psychoanalytic technique were introduced by several of Freud's colleagues who parted company with him. Analytic psychology, devised by Carl Jung, placed less emphasis on free association and more on the interpretation of dreams and fantasies. Special importance was given to the collective unconscious, a reservoir of shared unconscious wisdom and ancestral experience that entered consciousness only in symbolic form to influence thought and behaviour. Jungian analysts sought clues to their patients' problems in the archetypal nature of myths, stories, and dreams. Individual psychology, devised by Alfred Adler, emphasized the importance of the individual's drive toward power and of his unconscious feelings of inferiority. The therapist was concerned with the patient's compensations for his inferiority, as well as with his social relationships.

Development of physical and pharmacological treatments.

During the early decades of the 20th century the principal approaches to the treatment of mental disorders were psychoanalytically derived psychotherapies, used to treat people with neuroses, and custodial care in mental hospitals, for those with psychoses. But beginning in the 1930s these methods began to be supplemented by physical approaches using drugs, electroconvulsive therapy, and surgery. The first successful physical treatment in psychiatry was the induction of malaria in patients with a fatal form of neurosyphilis called general paresis. The malarial treatment stemmed from the observation that some psychotic patients improved during febrile illnesses. In 1933 the Polish psychiatrist Manfred Sakel reported the treatment of schizophrenia by repeated insulin-induced comas. (Neither of these treatments is in use today.) The treatment of schizophrenia by convulsions, originally induced by the injection of camphor, was reported in 1935 by the psychiatrist Ladislaus Joseph von Meduna in Budapest. An improvement in this approach was the induction of convulsions by the passage of an electrical current through the brain, a technique introduced by the Italian psychiatrists Ugo Cerletti and Lucio Bini in 1938. Electroconvulsive treatment was more successful in alleviating states of severe depression than in treating schizophrenia.

Psychosurgery, or surgery performed to treat mental illness, was introduced by the Portuguese neurologist António Egas Moniz in the 1930s. The operation Moniz originated, leucotomy, or lobotomy, was widely performed during the next two decades in the treatment of schizophrenia, intractable depression, severe anxiety, and severe obsessional states. The procedure was later abandoned, however, largely because its therapeutic effects could be better obtained by the use of newly developed drugs.

The decades after World War II were marked by the first safe and effective applications of drugs in the treatment of mental disorders. Prior to the 1950s such sedative compounds as bromides and barbiturates had been used to quiet or sedate patients, but these drugs were general in their effect and did not target the specific symptoms of mood disturbances or psychotic disorders. Many of the drugs that subsequently proved effective in treating such conditions were recognized

serendipitously; i.e., when researchers administered them to patients just to see what would happen, or when they were administered to treat one mental condition and were instead found to be helpful in alleviating the symptoms of an entirely different condition.

The first effective pharmacological treatment of psychosis was the treatment of mania with lithium, introduced by the Australian psychiatrist J.F.J. Cade in 1949. Lithium, however, excited little interest until its dramatic effectiveness in the maintenance treatment of bipolar affective disorder was reported in the mid-1960s. Chlorpromazine, the first of a long series of highly successful antipsychotic drugs, was synthesized in France in 1950 during work on antihistamines. It was used in anesthesia before its antipsychotic and tranquilizing effects were reported in France in 1952. The first tricyclic antidepressant drug, imipramine, was originally designed as an antipsychotic drug and was investigated by the Swiss psychiatrist Roland Kuhn. He found it ineffective in treating schizophrenia but observed its antidepressant effect, which he reported in 1957. A drug used in the treatment of tuberculosis, iproniazid, was found to be effective as an antidepressant in the mid-1950s. It was the first monoamine oxidase inhibitor to be used in psychiatry. The first modern anxiety-relieving drug was meprobamate, which was originally introduced as a muscle relaxant. It was soon overtaken by the pharmacologically rather similar but clinically more effective chlordiazepoxide, which was synthesized in 1957 and marketed as Librium in 1960. This drug was the first of the extensively used benzodiazepines. These and other drugs had a revolutionary impact not only on psychiatry's ability to relieve the symptoms and suffering of people with a wide range of mental disorders but also on the institutional care of the mentally ill.

Deinstitutionalization.

Between about 1850 and 1950 there was a steady increase in the number of patients staying in mental hospitals. In England and Wales, for example, there were just over 7,000 such patients in 1850, nearly 120,000 in 1930, and nearly 150,000

in 1954. Then a steady decline in the number of occupied beds began, reaching just over 100,000 in 1970 and 75,000 in 1980, a decrease of almost 50 percent. The same process began in the United States in 1955 but continued at a more rapid rate. The decrease, from just under 560,000 in 1955 to just over 130,000 in 1980, was one of more than 75 percent. In both countries it became official policy to replace mental hospital treatment with community care, involving district general hospital psychiatric units in Britain and local mental health centres in the United States. This dramatic change can be partly attributed to the introduction of antipsychotic drugs, which quieted many psychotic patients and drastically changed the atmosphere of mental hospital wards for the better. With the recovery of lucidity and calmness, many psychotic patients could return to their homes and live at least a partially normal existence instead of spending their lives sequestered in mental hospitals. The wholesale release of mental patients into the community was not without problems, however, since many areas lacked the facilities to support and maintain such patients, many of whom thus received inadequate care.

Development of behaviour therapy.

In the 1950s and '60s a new type of therapy, called behaviour therapy, was developed. In contrast to the existing psychotherapies, its techniques were based on theories of learning derived from research on classical conditioning carried out by Ivan Pavlov and others, and from the work of such American behaviourists as John B. Watson and B.F. Skinner. Behavioral therapy developed when the theoretical principles that were originally developed from experiments with animals were applied to the treatment of patients.

In 1920 Watson experimentally induced a phobia of rats in a small boy, and in 1924 Mary Cover Jones reported the extinction of phobias in children by gradual desensitization. Modern behaviour therapy began with the description by the South African psychiatrist Joseph Wolpe of his technique of systematically desensitizing a patient with phobias, beginning by exposing him to the least and gradually progressing to the most feared object or situation. Behavioral therapies were more

quickly adopted in Europe than in the United States, where psychoanalytic precepts had exercised a particular dominance over psychiatry, but by the 1980s behavioral therapies were also well established in the United States.

The mental health profession in the late 20th century.

There has been a great increase in the number of mental health professionals since World War II. In the United States the number of psychiatrists was 3,000 in 1939 but had increased to more than 25,000 by the early 1970s. Nonmedical mental health professionals have also increased in number and have achieved increasing independence from medical control, acquiring new roles in the process. Clinical psychologists, for instance, who were at one time largely confined to carrying out psychometric tests at the doctor's request, have become increasingly concerned with psychotherapy and behaviour therapy. Psychiatric social workers are no longer confined to casework with individual patients or their relatives but have also become psychotherapists and play prominent roles in mental health centres. There are new roles for nurses, including behaviour therapy and the management of chronic mental illness in the community. The greatest beneficiaries of this expansion of mental health professionals have been patients with neurotic and other less severe disorders.

Psychotherapy retains a major role in the mental health profession, and since the development of psychoanalysis the varieties of psychotherapy have increased and multiplied to the extent that a 1980 handbook listed "more than 250 different therapies in use today." The repertoire of drugs used in the treatment of mental illness has continued to grow as new drugs are developed or new applications of existing ones are discovered, and research on the biochemical and genetic causes of mental disease continues to make gradual headway in explicating the causes of various disorders. The triad of psychotherapy, drugs, and behaviour therapy afford an unprecedented array of approaches, techniques, and procedures for alleviating symptoms or curing people altogether of mental disorders.

PHYSIOLOGICAL TREATMENTS

Pharmacological treatments.

Antipsychotic agents. Antipsychotic drugs, which are also known as neuroleptics and major tranquilizers, belong to several different chemical groups but are similar in their therapeutic effects. They have a calming effect that is valuable in the relief of agitation, excitement, and violent behaviour in psychotic patients. The drugs are quite successful in reducing the symptoms of schizophrenia, mania, and delirium, and they are used in combination with antidepressants to treat psychotic depression. The drugs suppress hallucinations and delusions, alleviate disordered or disorganized thinking, improve the patient's lucidity, and generally make him more receptive to psychotherapy. Patients who have previously been agitated, intractable, or grossly delusional become noticeably calmer, quieter, and more rational when maintained on these drugs. The drugs have enabled many episodically psychotic patients to have shorter stays in hospitals, and many other patients who would have been permanently confined to institutions are able to live in the outside world because of the drugs. The antipsychotics differ in their unwanted effects: some are more likely to make the patient drowsy; some to alter blood pressure or heart rate; and some to cause tremor or slowness of movement. Some of the most widely used antipsychotic drugs are shown in Table 1.

In the treatment of schizophrenia, antipsychotic drugs partially or completely control such symptoms as delusions and hallucinations. They also protect the patient who has recovered from an acute episode from suffering a relapse. Unfortunately, the drugs have less of an effect on such symptoms as social withdrawal, apathy, blunted emotional capacity, and the other psychological deficits characteristic of the chronic stage of the illness.

No single drug seems to be outstanding in the treatment of schizophrenia. In an individual patient, one drug may be preferred to another because it produces less severe unwanted effects, and the dose of any one drug needed to produce a therapeutic effect varies widely from patient to patient. Because of these individual

differences it is common for psychiatrists to substitute a drug of a different chemical group when one drug has been shown to be ineffective despite its use in adequate dosage for several weeks.

In an acute psychotic episode a drug such as chlorpromazine, trifluoperazine, or haloperidol usually has a calming effect within a day or two. The control of psychotic symptoms such as hallucinations or disordered thinking may take weeks. The appropriate dosage has to be determined for each patient by cautiously increasing the dose until a therapeutic effect is achieved without unacceptable side effects.

Antipsychotic drugs are thought to work by blocking dopamine receptors in the brain. Dopamine is a neurotransmitter; i.e., a chemical messenger produced by certain nerve cells that influence the function of other nerve cells by interacting with receptors in their cell membranes. Since schizophrenia may be caused by either the excessive release of or an increased sensitivity to dopamine in the brain, the effects of antipsychotic drugs may be due to their ability to block or inhibit dopamine transmission.

Dopamine receptor blockade is responsible for the drug's main unwanted side effects, which are termed extrapyramidal symptoms (EPS). These resemble the symptoms of Parkinson's disease and include tremor (shakiness) of the limbs; bradykinesia - slowness of movement with loss of facial expression, absence of arm-swinging during walking, and a general muscular rigidity; dystonia - sudden, sustained contraction of muscle groups causing abnormal postures; akathisia - a subjective feeling of restlessness leading to an inability to keep still; and tardive dyskinesia - involuntary movements, particularly involving the lips and tongue. Most extrapyramidal symptoms disappear when the drug is withdrawn. Tardive dyskinesia occurs late in the drug treatment and in about half of the cases persists even after the drug is no longer used. There is no satisfactory treatment.

Antianxiety agents.

The drugs most commonly used in the treatment of anxiety are the benzodiazepines, or minor tranquilizers, which have replaced the barbiturates because of their vastly greater safety. The advantage of benzodiazepines is that they calm the patient without the marked sleep-inducing effects of barbiturates, so that a degree of wakeful alertness is maintained and the individual can carry on his daily activities. A large number of benzodiazepines have been marketed, and the more common ones are listed in Table 2. They differ from each other in duration of action rather than effectiveness. Smaller doses have a calming effect and alleviate both the physical and psychological symptoms of anxiety. Larger doses induce sleep, and some benzodiazepines are marketed as hypnotics. The benzodiazepines have become among the most widely prescribed drugs in the developed world, and controversy has arisen over their excessive use by the public.

The side effects of these drugs are usually few - most often drowsiness and unsteadiness. The drugs themselves are not lethal even in very large overdoses, but they increase the sedative effects of alcohol and other drugs. The benzodiazepines are basically intended for short- or medium-term use, since the body develops a tolerance to them that reduces their effectiveness and necessitates the use of progressively larger doses. Dependence on them also occurs, even in moderate dosages, and withdrawal symptoms have been observed in those who have used the drugs for only four to six weeks. In patients who have taken a benzodiazepine for many months or longer, withdrawal symptoms occur in between 15 and 40 percent of the cases and may take weeks or months to subside.

The withdrawal symptoms are of three kinds. Such severe symptoms as delirium or convulsions are rare. Frequently the withdrawal symptoms represent a renewal or increase of the anxiety itself. Many patients also experience other symptoms such as hypersensitivity to noise and light, as well as muscle twitching. As a result, many long-term users continue to take the drug not because of persistent anxiety but because the withdrawal symptoms are too unpleasant.

Because of the danger of dependence, benzodiazepines should be taken in the lowest possible dose for no more than a few weeks. For longer periods they should be taken intermittently, and only when the anxiety is severe.

Benzodiazepines act on specialized receptors in the brain that are adjacent to receptors for a neurotransmitter called gamma-aminobutyric acid (GABA), which inhibits anxiety. It is possible that the interaction of benzodiazepines with these receptors facilitates the inhibitory (anxiety-suppressing) action of GABA within the brain.

Antidepressant agents.

Many patients suffering from depressive illness gain symptomatic relief from treatment with one of the tricyclic drugs (so called because of their three-ringed chemical structure) or one of the newer drugs with similar properties. Some of the most widely used antidepressant drugs are shown in Table 3.

Patients with melancholia who are not delusional benefit the most from the antidepressants. The drug is given in reduced dosage for the first few days to lessen the severity of any side effects, which are often more severe in the first days of treatment, and the drug is then given in full dosage for at least three to four weeks. A sedative effect may be apparent within the first two to three days, but it usually takes two to three weeks for the drug to significantly improve the patient's depressed mood. If there is no improvement after four weeks of adequate dosage, the drug should be discontinued over the course of a few days. There is no convincing evidence that any one tricylic antidepressant is superior to the others in therapeutic effectiveness. If a sedative antidepressant produces troublesome daytime drowsiness, it should be replaced by a less sedative drug. If the side effects of a classical tricyclic are severe, it can be replaced by lofepramine, a newer drug with fewer unwanted effects, or with one of the tricyclic-like drugs.

Successful treatment with such drugs relieves all the symptoms of depressive illness, including disturbances of sleep and appetite, loss of sexual desire, and decreased energy, interest, and concentration. Once a good response has been

achieved, the drug should be continued in reduced dosage for a further six months. Research has shown that such maintenance treatment greatly reduces the risk of relapse during that time.

It is widely theorized that depression is partly caused by reduced quantities or reduced activity of the monoamine neurotransmitters serotonin and norepinephrine. Tricyclics are thought to act by inhibiting the body's physiological inactivation of the monoamine neurotransmitters in the brain. This results in the buildup or accumulation of the neurotransmitters there and allows them to remain in contact longer with their receptors, changes that may be important in elevating mood.

The side effects of these drugs are mostly due to their interference with the function of the autonomic nervous system. The side effects include dryness of the mouth, blurred vision, constipation, and, particularly in older men, difficulty in passing urine. Weight gain can be a distressing side effect in patients taking a tricyclic for a long period. Such patients often report an increase in appetite for carbohydrate-rich foods. In elderly patients, in whom lower doses should be used, antidepressants can cause delirium. The classic tricyclics interfere with conduction in heart muscle, and so they are best avoided in patients with heart disease. They are also dangerous if a patient takes a large overdose of them. Drug interactions occur with tricyclics, the most important being their interference with the action of certain drugs used in the treatment of high blood pressure.

Monoamine oxidase inhibitors (MAOIs), of which phenelzine is the best researched, are generally ineffective in patients with melancholia or depressive delusions but are more effective in patients whose depressed mood can be temporarily lightened by a change in the environment. They are used in the same way as tricyclics, with a low initial dose increased after a few days. Their antidepressant effect is also delayed, and treatment needs to continue for four to six weeks before its effectiveness can be assessed.

As their name implies, the drugs interfere with the action of monoamine oxidase, an enzyme involved in the breakdown of norepinephrine and serotonin. As a result,

these neurotransmitters accumulate within nerve cells and presumably leak out onto receptors.

The side effects of these drugs include daytime drowsiness, difficulty in getting to sleep, and a fall in blood pressure when rising to one's feet. The MAOIs interact dangerously with various other drugs, including narcotics and some over-the-counter drugs used in treating colds. Patients taking an MAOI must avoid certain foods containing tyramine or other naturally occurring amines, which can cause a severe rise in blood pressure leading to headaches and even to intracranial bleeding. Tyramine occurs in several foodstuffs, of which the most important are cheese, Chianti wine, and well-cured meats. The MAOI drugs are safe in normal dosages but are dangerous in overdosage.

Certain antidepressants seem to be effective in treating other mental disorders. This is true of imipramine in some cases of panic disorder, monoamine oxidase inhibitors in agoraphobia, and clomipramine in agoraphobia and in some cases of obsessive-compulsive disorder.

Mood-stabilizing drugs.

Lithium, usually administered as its carbonate in several small doses per day, is effective in the treatment of an episode of mania. It can drastically reduce the elation, overexcitement, grandiosity, paranoia, irritability, and flights of ideas typical of people in the manic state. It has little or no effect for several days, however, and a therapeutic dose is rather close to a toxic dose. In severe episodes haloperidol or chlorpromazine may also be used. Lithium also has an antidepressant action in some patients with melancholia.

The most important use of lithium is in the maintenance treatment of patients with bipolar affective disorder (manic-depressive illness) or with recurrent depression. When given while the patient is well, lithium may prevent further mood swings, or it may reduce either their frequency or their severity. Its mode of action is unknown. Treatment begins with a small dose that is gradually increased until a specified concentration of lithium in the blood is reached. Blood tests to determine

this are carried out weekly in the early stages of treatment and later every two to three months. It may take as long as a year for lithium to become fully effective.

The toxic effects of lithium, which usually occur when there are high concentrations of it in the blood, include drowsiness, coarse tremors, vomiting, diarrhea, incoordination of movement, and, with still higher blood concentrations, convulsions, coma, and death. At therapeutic blood concentrations, lithium's side effects include fine tremors (which can be alleviated by propranolol), weight gain, passing increased amounts of urine with consequent increased thirst, and reduced thyroid function.

Carbamazepine, an anticonvulsant drug with a chemical composition similar to that of the tricyclic antidepressants, has been shown to be effective in the treatment of mania and in the maintenance treatment of bipolar affective disorder. It may be combined with lithium in bipolar affective patients who fail to respond to either drug alone.

Electroconvulsive treatment.

In electroconvulsive treatment (ECT) a convulsion is produced in a person by passing an electric current through his brain. The duration of the convulsive activity in the brain appears to determine its therapeutic effects, while the intensity of the electrical stimulus plays a role in determining its unwanted side effects, particularly the short-term memory impairment in the patient immediately after treatment. Several controlled trials have shown that ECT is effective in treating patients suffering from a depressive illness with melancholia.

Prior to the administration of ECT, the patient is given an intravenous injection of a short-acting anaesthetic to put him to sleep and then is given an injection of a muscle-relaxant in order to reduce the force of his muscular contractions during the convulsion. The electrical current is then applied to the brain. In bilateral ECT this is done by applying an electrode to each side of the head; in unilateral ECT both electrodes are placed over the nondominant cerebral hemisphere - i.e., the right side of the head in a right-handed person. Unilateral ECT produces noticeably less

confusion and memory impairment in patients, but more treatments may be needed. Patients recover consciousness rapidly after the treatment but may be confused and may experience a mild headache for an hour or two.

ECT treatments are normally given two or three times a week in the treatment of depressed patients. Once a program of ECT has been successfully completed, maintenance treatment with a tricyclic drug such as imipramine for the next few months significantly decreases the patient's risk of relapse.

Whenever rapid improvement is important, ECT is preferred to the treatment of depression with drugs. This is so in cases of severe depression when the patient's life is endangered because of refusal of food and fluids or because of serious risk of suicide, as well as in cases of depression with psychosis after childbirth, when it is desirable to reunite the mother and baby as soon as possible. ECT is also used in treating patients with melancholia who have failed to respond to adequate dosages of antidepressants, and to treat elderly depressed patients who are unable to tolerate antidepressant drugs.

The number of electroconvulsive treatments required to treat depression is usually between four and eight, with more needed by some elderly patients. Some patients improve after the first treatment, others only after several. The mode of action of ECT is not understood. It has been shown that electrically induced convulsions in animals alter the sensitivity of norepinephrine and serotonin receptors in the brain. Conceivably this action could alter depressed mood in human patients.

The chief unwanted effect of ECT is impairment of memory. Some patients report memory gaps covering the period just before treatment, but others lose memories from several years before treatment. Many patients have memory difficulties for a few days or even a few weeks after completion of the treatment so that they forget appointments, phone numbers, and the like. These difficulties are transient and disappear rapidly in the vast majority of patients. Occasionally, however, patients complain of permanent memory impairment after ECT; these are almost always patients who did not recover from their depression as a result of the treatment.

Psychosurgery.

Psychosurgery is the destruction of groups of nerve cells or nerve fibres in the brain by surgical techniques in an attempt to relieve psychiatric symptoms that are not due to structural brain disease. The removal of a brain tumour that is causing psychiatric symptoms is not an example of psychosurgery.

The classical technique of bilateral prefrontal leucotomy (lobotomy) is no longer performed because of its frequent undesirable effects on physical and mental health, in particular the development of epilepsy and the appearance of permanent, undesirable changes in personality. The latter included increased apathy and passivity, lack of initiative, and a generally decreased depth and intensity of the person's emotional responses to life. In its heyday the operation was used to quiet chronically tense, delusional, agitated, or violent psychotic patients. Stereotaxic surgical techniques have been developed that enable the surgeon to produce small areas of nerve cell or fibre destruction by means of metal probes inserted into accurately located parts of the brain. The nerve tissue is then destroyed by the implantation of a radioactive substance (usually yttrium) or by the application of heat or cold.

The proponents of psychosurgery claim that it is effective in treating some patients with severe and intractable depression and with anxiety or obsessional neuroses and that it may improve the behaviour of abnormally aggressive patients. There is no compelling evidence to support these claims, and many of the therapeutic effects that were claimed for psychosurgery by its adherents are in fact now attainable by the use of antipsychotic and antidepressant drugs. Many physicians believe that psychosurgery is never justified. Others accept that it has a very small part to play in psychiatric treatment when the prolonged use of other forms of treatment has been unsuccessful and the patient is chronically and severely distressed or tormented by his symptoms. Whereas ECT is a routine treatment in certain specified conditions, psychosurgery is at best a last resort.

THE PSYCHOTHERAPIES

Psychotherapy implies the treatment of mental discomfort, dysfunction, or disease by psychological means by a trained therapist who adheres to a particular theory of both symptom causation and symptom relief. The American psychotherapist Jerome Frank has classified psychotherapies into religio-magical and empirico-scientific forms. The former depend on the shared beliefs of the therapist and client in magic, spirits, or other supernatural processes or powers. This article is concerned, however, with the latter forms of psychotherapy, which have been developed by modern medicine and which are carried out by a member of one of the mental health professions such as a psychiatrist or a clinical psychologist.

It is usual to contrast two main forms of psychotherapy, dynamic and behavioral. They are conceptually different; behaviour therapy concentrates on alleviating a patient's overt symptoms, which are attributed to faulty learning, while dynamic therapy concentrates on understanding the meaning of symptoms and understanding the emotional conflicts within the patient that may be causing those symptoms. In their pure forms the two approaches are very different, but in practice many therapists use elements of both.

Dynamic psychotherapies.

There are many variants of dynamic psychotherapy, all of which ultimately derive from the basic precepts of psychoanalysis. The fundamental approach of most dynamic psychotherapies can be traced to three basic theoretical principles or assertions: (1) Human behaviour is prompted chiefly by emotional considerations, but insight and self-understanding are necessary to modify and control such behaviour and its underlying aims; (2) A significant proportion of human emotion is not normally accessible to one's personal awareness or introspection, being rooted in the unconscious, those portions of the mind beneath the level of consciousness; (3) Any process that makes available to a person's conscious awareness the true significance of emotional conflicts and tensions that were hitherto held in the unconscious will thereby produce heightened awareness and increased stability and emotional control. The classic dynamic psychotherapies are

relatively intensive talking treatments that are aimed at providing the person with insight into his conscious and unconscious mental processes, with the goal of enabling him ultimately to achieve a better understanding of himself.

Dynamic psychotherapy attempts to enhance the patient's personality growth as well as to alleviate his symptoms. The main therapeutic forces are activated in the relationship between patient and therapist and depend both upon the empathy, understanding, integrity, and concern demonstrated by the therapist and upon the motivation, intelligence, and capacity for achieving insight manifested by the patient. The attainment of a therapeutic alliance - i.e., a working relationship between patient and therapist that is based on mutual respect, trust, and confidence - provides the context in which the patient's problems can be worked through and resolved. Several of the most important forms are treated below.

Psychoanalytic psychotherapy.

Classical psychoanalysis is the most demanding of all the psychotherapies in terms of time, cost, and effort. It is conducted with the patient lying on a couch and with the analyst seated out of his sight but close enough to hear what the patient says. The treatment sessions last 50 minutes and are usually held four or five times a week for at least three years. The primary technique used in psychoanalysis and in other dynamic psychotherapies to enable unconscious material to enter the patient's consciousness is that of "free association." In free association, according to Freud, the patient "is to tell us not only what he can say intentionally and willingly, what will give him relief like a confession, but everything else as well that his self-observation yields him, everything that comes into his head, even if it is disagreeable for him to say it, even if it seems to him unimportant or actually nonsensical." Such a procedure is rendered difficult, first, because for a person to speak of his innermost (and often socially unacceptable) thoughts is a departure from years of practice in which he has selected what he has said to others. Free association is also difficult because the patient resists remembering repressed experiences or feelings that are connected with intense or conflicting emotions that

have never been finally resolved or settled. Such repressed emotions or memories usually revolve around the patient's important personal relationships and his innermost feelings of self, and the release or recollection of such emotions in the course of treatment can be itself intensely disturbing.

Attentive listening and "empathy" on the part of the therapist allows the patient to express thoughts and feelings that in turn permit the uncovering of his underlying emotional conflicts. In the course of treatment, the patient often seeks to project (attribute to something other than himself) the disturbing emotions he feels in the process of recollection and free association, and the person who is almost invariably selected for the focus of such projection is the psychoanalyst; that is, the patient is likely to blame his emotional distress on the analyst. In this way, the patient comes to feel love or hatred, dependence or rebellion, and rivalry or rejection toward the analyst. These are the same attitudes the patient has felt but has never consciously acknowledged toward his parents or other people with whom he shared important relations earlier in life. The patient's projection onto the therapist of these feelings and behaviours that originated in his earlier relationships is called the transference. To facilitate the development of the transference, the analyst endeavours to maintain a neutral stance toward the patient in order to serve as a "blank screen" onto which the patient can project his inner feelings. The analyst's handling of the transference situation is of vital importance in psychoanalysis or, indeed, in any form of dynamic psychotherapy. It is through the transference that the patient discovers the nature of his unconscious feelings and then becomes able to acknowledge them. Once this has been done, he often finds himself able to regard them in a far more dispassionate and tolerant light and often feels himself liberated from their influence upon his future behaviour.

A major therapeutic tool in the course of treatment is interpretation. This technique helps the patient to become aware of previously repressed aspects of his emotional conflicts and to uncover the meaning of uncomfortable feelings evoked by the transference. Interpretation, in turn, is used to determine the underlying psychological meaning of the patient's dreams, which are held to have a hidden or

latent content that symbolize and indirectly express aspects of the patient's emotional conflicts.

Individual dynamic psychotherapy.

Although psychoanalysis has had a profound influence, particularly on American psychiatry, that influence waned significantly during the 1970s and '80s. Fewer patients now enter psychoanalysis, and many analysts carry out short-term individual dynamic psychotherapy. This form of therapy is much more readily available and usually requires 50 minutes a week for six to 18 months. The aim of treatment, as in psychoanalysis, is to increase the patient's insight (understanding of himself), to relieve his symptoms, and to improve his psychological functioning. Suitable patients include those with a wide range of neurotic disorders and personal or social problems who wish to change and who are able to view their problems in psychological terms.

As in psychoanalysis, the patient learns to trust the therapist and becomes able to talk candidly and honestly about his most intimate thoughts and feelings. The treatment setting is less formal than in psychoanalysis, with the therapist and patient seated so that eye contact can be achieved if desired.

Treatment techniques include free association and the use of interpretation by the therapist to analyze the transference, the patient's unconscious defense mechanisms, and his dreams. The therapist may ask the patient to clarify or enlarge on some point on which the therapist is not clear if this seems important in the development of the patient's symptoms. The therapist directs the patient's attention to important links, of which he seems unaware, between the present and the past, between his emotional responses to the therapist and to people important to him, and so on. The therapist may challenge the patient with the likely consequences of his resistant or maladaptive behaviour and stress instead the importance of confronting and trying to resolve his psychological difficulties.

Brief focal psychotherapy.

This is a form of short-term dynamic therapy in which a time limit to the duration of the therapy is often agreed upon with the patient at the outset. Sessions lasting 30 to 60 minutes are held weekly for, typically, five to 15 weeks. At the beginning of treatment the therapist helps identify the patient's problem or problems, and these are made the focus of the treatment. The problem should be an important source of distress to the patient and it should be modifiable within the time limit. The therapist is more active, directive, and confrontational than in long-term dynamic therapy and ensures that the patient keeps to the focus of treatment and is not diverted by subsidiary problems or concerns. Some therapists deliberately aim to produce considerable emotional arousal in the patient during each session as a way to activate or highlight specific problems. Research suggests that brief therapy can produce as good results as long-term therapy, and more quickly.

Group psychotherapy.

Many types of psychological treatment may be provided for groups of patients with psychiatric disorders. This is true, for example, of relaxation training and anxiety-management training. There are also self-help groups, of which Alcoholics Anonymous is perhaps the best known. A considerable number of group experiences have been devised for people who are not suffering from any psychiatric disorder; encounter groups are a well-known example. This discussion, however, is concerned with long-term dynamic group therapy, in which six to 10 psychiatric patients meet with a trained group therapist, or sometimes two therapists, for 60 to 90 minutes a week for up to 18 months. Often the group is closed, i.e., confined to the original group membership, even if one or more members drop out before the treatment ends. In an open group patients who have stopped attending, whether by default or because of the relief of symptoms, are replaced by new members.

The types of mental disorders considered suitable for group therapy are much the same as those suitable for individual therapy. Again the patient must want to

change and must be psychologically minded. In addition, it is important that he not consider group therapy as a poor second to individual therapy.

There are many varieties of dynamic group therapy, and they differ in their theoretical background and technique. The influential model of the American psychiatrist Irvin D. Yalom provides a good example of such therapies, however. The therapist continually encourages the patients to direct their attention to the personal interactions occurring within the group rather than to what happened in the past to individual members or what is currently happening outside the group, although both of these areas may be considered when they are relevant. The therapist regularly draws attention to what is happening among members of the group as they learn more about themselves and test out different ways of behaving with one another. The goal in group therapy is to create a climate in which the participants can shed their inhibitions. When the members come to trust one another, they are able to provide feedback and to respond to other group members in ways that might not be possible in ordinary social interactions owing to the constraints of social conventions.

Several factors appear to be important in group therapy. The most important is group cohesion, which gives the patient a feeling of belonging, identification, and security and enables him to be frank and open and to take risks without the danger of rejection. Universality refers to the patient's realization that he is not unique, that all the other group members have problems, some of them similar to his. Optimism about what can be achieved in the group, fostered by the perception of change in others, combats demoralization. Guidance, the giving of advice and explanation, is important in the early meetings of the group and is largely a function of the therapist. What has been called vicarious learning later becomes more important; the patient observes how other group members evolve solutions to common problems and emulates desirable qualities he sees in fellow members. Catharsis, or the release of highly charged emotion, occurs within the group setting and is helpful provided that the patient can come to understand it and appreciate its

significance. Another factor that is helpful in improving self-esteem is altruism, the opportunity to give assistance to another group member.

Family therapy.

Family therapists view the family as the "patient" or "client" and as more than the sum of its members. The family as a focus for treatment usually comprises the members who live under the same roof, sometimes supplemented by relatives who live elsewhere or by nonrelatives who share the family home. Therapy with couples - marital therapy - may be considered as a special type of family therapy. Family therapy may be appropriate when the person referred for treatment has symptoms clearly related to such disturbances in family function as marital discord, distorted family roles, and parent-child conflict, or when the family as a unit asks for help. It is not appropriate when the patient has a severe disorder needing specific treatment in its own right.

The many theoretical approaches include psychoanalytic, systems-theory, and behavioral models. The analyst is concerned with the family's past as the cause of the present; he pays attention to psychodynamic aspects of the individual members as well as of the family unit; and he makes numerous interpretations and aims at increasing the insight of the members. The systems therapist, on the other hand, is interested in the present rather than the past and is often not concerned with promoting insight but rather with changing the family system, perhaps by altering the implicit and fixed rules under which it functions so that it can do so more effectively. The behaviour therapist is concerned with behaviour patterns and with pinpointing the types of reinforcement that maintain behaviour that other family members regard as undesirable. Members specify the changes in behaviour that they wish to see in one another, and strategies are devised to reinforce the desired behaviours. This approach has been shown to be effective in work with couples, when one partner promises some particular change provided that the other reciprocates.

Treatment sessions in family therapy are rarely held more often than once a week and often only once every three or four weeks. Termination commonly occurs when the therapist considers that treatment has succeeded - or failed irretrievably - or when the family firmly decides to withdraw from treatment. There seems no doubt that family therapy can produce marked change within a family.

Behavioral psychotherapy.

This is an approach to the treatment of mental disorders that uses a variety of methods based upon principles derived from experimental psychology, mainly that of learning theory. In the words of Joseph Wolpe, "behavior therapy, or conditioning therapy, is the use of experimentally established principles of learning for the purpose of changing unadaptive behavior. Unadaptive habits are weakened and eliminated; adaptive habits are initiated and strengthened" (The Practice of Behavior Therapy, 1973).

In the treatment of phobias, behaviour therapy seeks to modify and eliminate the avoidance response that the patient manifests when confronted with a phobic object or situation. This is crucial because, since the patient's avoidance of the anxiety-producing situation does indeed reduce his anxiety, his conditioned association of the phobic situation with the experience of anxiety remains unchallenged and frequently persists. Behaviour therapy interrupts this self-reinforcing pattern of avoidance behaviour by presenting the feared situation to the patient in a controlled manner such that it eventually ceases to produce anxiety; the strong associative link that has been built up within him between the feared situation, the experience of anxiety, and his subsequent avoidance behavior will thus be broken down and replaced by a less maladaptive set of responses.

The behavioral therapist is concerned with the forces and mechanisms hat perpetuate the patient's present symptoms or abnormal behaviours, not with experiences in the past that may have caused them nor with any postulated intrapsychic conflict. Behavioral therapy focuses on observable phenomena; i.e.,

what is done and what is said, rather than on what must be inferred (unconscious motives and processes and symbolic meanings).

The behavioral therapist carries out a detailed analysis of the patient's behaviour problems, paying particular attention to the circumstances in which they occur, to the patient's attempts to cope with his symptoms, and to his wish for change. The goals of treatment are precisely defined and usually do not include aims such as personal growth or personality change. The relationship between patient and therapist is sometimes said to be unimportant in behaviour therapy, and it is quite true that a patient may successfully follow a behavioral therapeutic program that he learns from a book or a personal computer. Nevertheless, a patient is more likely to complete an arduous program if he trusts and respects the therapist.

Behaviour therapy has become the preferred treatment for phobic states and for some obsessive-compulsive disorders, and it is effective in many cases of sexual dysfunction and deviation. It also has an important role in the rehabilitation of patients with chronic, disabling disorders.

Treatment of neurotic disorders.

The essence of the treatment of phobias is the controlled exposure of the patient to the objects or situations that he fears. Behaviour therapy tries to eliminate the phobia by teaching the patient how to face those situations that clearly trigger his discomfort so that he can learn to tolerate them. The exposure of the patient to the feared situation can be gradual (sometimes called desensitization) or rapid (sometimes known as flooding). Contrary to popular belief, the anxiety that is produced during exposure is not usually harmful. Even if severe panic does strike the sufferer, it will gradually evaporate and will be less likely to return in the future. Effective exposure treatments developed only as therapists learned to endure the phobic anxiety exhibited by their patients and when the therapists grew secure in the knowledge that such anxiety is much more likely to lead to the patients' improvement than to be harmful. The important point in this therapy is to persevere until the phobic anxiety starts to lessen and to be prepared to go on until

it does. In general, the more rapidly and directly the worst fears are embraced by the patient, the more quickly the phobic terror fades to a tolerable mild tension.

In the technique of desensitization, the patient is first taught how to practice muscular relaxation. He then reviews the situations of which he is afraid and lists them in order of increasing fearfulness, called a "hierarchy." Finally, the patient faces the various fear-producing situations in ascending order by means of vividly imagining them, countering any anxiety he feels by using relaxation techniques. This treatment is prolonged, and its use is restricted to feared situations that patients cannot regularly confront in real life, such as fear of lightning.

One of the most common phobic disorders treated by exposure techniques is agoraphobia (fear of open or public places). The patient, who is commonly a young woman, is encouraged to practice exposure daily, staying in a phobic situation for at least an hour, so that anxiety has time to reach a peak and then subside. The patient must be determined to get the better of the fears and not to run away from them. Agoraphobics cannot simply enter a dreaded crowded store, feel the familiar surge of panic, and rush out again. They must devote a full afternoon to a shopping trip. When panic strikes, the patient can sit in a corner of the store and ride out the terror. When feeling better, the patient can continue shopping. Persistence and patience are essential to conquering phobias in this way.

There is considerable evidence that exposure techniques work in most cases. Even phobias present for as long as 20 years can be overcome in a treatment program requiring no more than three to 15 hours of therapist time per patient. There is also considerable evidence that many phobics can treat themselves perfectly adequately by self-exposure without a therapist, using carefully devised self-help manuals.

Some patients with obsessive-compulsive disorders can also be helped by behaviour therapy. Several different techniques may need to be employed. For instance, a patient with an obsessional fear of contamination is treated by exposure, being taught to "soil" his hands with dirt and then to avoid washing them for longer and longer periods. Anxiety-management training enables the patient to withstand the anxiety he feels during the period of exposure.

This and other techniques have been shown to be effective in the treatment of compulsive rituals, with improvement occurring in more than two-thirds of patients. There is also a reduction in the frequency and intensity of obsessional thoughts that accompany the rituals. The treatment of obsessional thoughts that occur alone is much less satisfactory, however.

Other therapies.

Many other types of psychotherapy have been developed in the second half of the 20th century, each with its own emphasis on symptom causation and its own particular approach to treatment. Many of these therapies use classical dynamic and behavioral models in modified forms, and they may also stress the understanding and modification of cognition and the ways in which people "process" their experiences, moods, and emotions. Among these relatively recent psychotherapies are client-centred psychotherapy, developed by the American psychologist Carl R. Rogers; transactional analysis, originated by the American psychiatrist Eric Berne; the interpersonal therapies developed by the American psychiatrists Adolf Meyer and Harry Stack Sullivan; cognitive therapy, developed by the American psychiatrist Aaron T. Beck; rational-emotive psychotherapy, developed by the American psychologist Albert Ellis; and Gestalt therapy, which stems from the work of the German psychiatrist Frederick S. (Fritz) Perls.

Another class of therapies consists of those used to care for psychotic patients, both those in hospitals and those who live in the community. Supportive psychotherapy consists of the long-term help of patients who are chronically handicapped by schizophrenia or other mental disorders. Such a therapy uses reassurance, guidance, and encouragement to help the patient cope with his disabilities and live as satisfactory a life as possible. Rehabilitation programs for chronic or episodically psychotic patients include drug maintenance; training in social skills that they may have lost while sick; occupational training to improve the patient's skills in cooking, shopping, and other domestic tasks; and industrial therapy, which usually offers the patient gainful employment under conditions of

minimal stress. Family therapy is sometimes used to help relatives learn to cope with a schizophrenic patient who has returned home from the hospital.

Community care for released schizophrenics or other psychotic patients must provide them with drug maintenance and a minimum of psychiatric monitoring; appropriate housing facilities; some type of employment; and training in such skills as using public transport, preparing their own food, and looking after their finances. Each patient should have a case manager, a professional worker who maintains contact and secures from governmental or social agencies the assistance that the patient needs. When provisions like these are not made, some formerly hospitalized patients stop taking their medicine and in effect drop out of the mental health care system, becoming unemployed and even homeless. This phenomenon became particularly evident in the United States and to a lesser extent in western Europe, when massive numbers of mental patients were released from hospitals during the 1950s and '60s after the effectiveness of antipsychotic drugs had been verified. These releases were also motivated by the concerns of civil libertarians over the abuse of patients' rights in keeping them committed to mental hospitals. However, the support network of community-based mental health clinics that would have been necessary to cope with the released patients was either inadequately established or nonexistent. The result was that many psychotic patients received inadequate outpatient care and supervision or encountered severe difficulties in obtaining housing or employment, becoming homeless wanderers in large urban areas.

ИНЖЕНЕРНО-ФИЗИЧЕСКИЙ ФАКУЛЬТЕТ ВЫСОКИХ ТЕХНОЛОГИЙ

Engineering

Introduction

Engineering is the professional art of applying science to the optimum conversion of the resources of nature to the uses of humankind. Engineering has been defined by the Engineers Council for Professional Development, in the United States, as the creative application of "scientific principles to design or develop structures, machines, apparatus, or manufacturing processes, or works utilizing them singly or in combination; or to construct or operate the same with full cognizance of their design; or to forecast their behaviour under specific operating conditions; all as respects an intended function, economics of operation and safety to life and property." The term engineering is sometimes more loosely defined, especially in Great Britain, as the manufacture or assembly of engines, machine tools, and machine parts.

The words engine and ingenious are derived from the same Latin root, ingenerare, which means "to create." The early English verb engine meant "to contrive." Thus the engines of war were devices such as catapults, floating bridges, and assault towers; their designer was the "engine-er," or military engineer. The counterpart of the military engineer was the civil engineer, who applied essentially the same knowledge and skills to designing buildings, streets, water supplies, sewage systems, and other projects.

Associated with engineering is a great body of special knowledge; preparation for professional practice involves extensive training in the application of that knowledge. Standards of engineering practice are maintained through the efforts of professional societies, usually organized on a national or regional basis, with each member acknowledging a responsibility to the public over and above responsibilities to his employer or to other members of his society.

The function of the scientist is to know, while that of the engineer is to do. The scientist adds to the store of verified, systematized knowledge of the physical

world; the engineer brings this knowledge to bear on practical problems. Engineering is based principally on physics, chemistry, and mathematics and their extensions into materials science, solid and fluid mechanics, thermodynamics, transfer and rate processes, and systems analysis.

Unlike the scientist, the engineer is not free to select the problem that interests him; he must solve problems as they arise; his solution must satisfy conflicting requirements. Usually efficiency costs money; safety adds to complexity; improved performance increases weight. The engineering solution is the optimum solution, the end result that, taking many factors into account, is most desirable. It may be the most reliable within a given weight limit, the simplest that will satisfy certain safety requirements, or the most efficient for a given cost. In many engineering problems the social costs are significant.

Engineers employ two types of natural resources - materials and energy. Materials are useful because of their properties: their strength, ease of fabrication, lightness, or durability; their ability to insulate or conduct; their chemical, electrical, or acoustical properties. Important sources of energy include fossil fuels (coal, petroleum, gas), wind, sunlight, falling water, and nuclear fission. Since most resources are limited, the engineer must concern himself with the continual development of new resources as well as the efficient utilization of existing ones.

Engineering as a profession

HISTORY OF ENGINEERING

The first engineer known by name and achievement is Imhotep, builder of the Step Pyramid at Saqqarah, Egypt, probably in about 2550 BC. Imhotep's successors - Egyptian, Persian, Greek, and Roman - carried civil engineering to remarkable heights on the basis of empirical methods aided by arithmetic, geometry, and a smattering of physical science. The Pharos (lighthouse) of Alexandria, Solomon's Temple in Jerusalem, the Colosseum in Rome, the Persian and Roman road systems, the Pont du Gard aqueduct in France, and many other large structures, some of which endure to this day, testify to their skill, imagination, and daring. Of

many treatises written by them, one in particular survives to provide a picture of engineering education and practice in classical times: Vitruvius' De architectura, published in Rome in the 1st century AD, a 10-volume work covering building materials, construction methods, hydraulics, measurement, and town planning.

In construction medieval European engineers carried technique, in the form of the Gothic arch and flying buttress, to a height unknown to the Romans. The sketchbook of the 13th-century French engineer Villard de Honnecourt reveals a wide knowledge of mathematics, geometry, natural and physical science, and draftsmanship.

In Asia, engineering had a separate but very similar development, with more and more sophisticated techniques of construction, hydraulics, and metallurgy helping to create advanced civilizations such as the Mongol empire, whose large, beautiful cities impressed Marco Polo in the 13th century.

Civil engineering emerged as a separate discipline in the 18th century, when the first professional societies and schools of engineering were founded. Civil engineers of the 19th century built structures of all kinds, designed water-supply and sanitation systems, laid out railroad and highway networks, and planned cities.

England and Scotland were the birthplace of mechanical engineering, as a derivation of the inventions of the Scottish engineer James Watt and the textile machinists of the Industrial Revolution. The development of the British machine-tool industry gave tremendous impetus to the study of mechanical engineering both in Britain and abroad.

The growth of knowledge of electricity - from Alessandro Volta's original electric cell of 1800 through the experiments of Michael Faraday and others, culminating in 1872 in the Gramme dynamo and electric motor (named after the Belgian Z.T. Gramme) - led to the development of electrical and electronics engineering. The electronics aspect became prominent through the work of such scientists as James Clerk Maxwell of Britain and Heinrich Hertz of Germany in the late 19th century. Major advances came with the development of the vacuum tube by Lee De Forest of the United States in the early 20th century and the invention of the transistor in

the mid-20th century. In the late 20th century electrical and electronics engineers outnumbered all others in the world.

Chemical engineering grew out of the 19th-century proliferation of industrial processes involving chemical reactions in metallurgy, food, textiles, and many other areas. By 1880 the use of chemicals in manufacturing had created an industry whose function was the mass production of chemicals. The design and operation of the plants of this industry became a function of the chemical engineer.

ENGINEERING FUNCTIONS

Problem solving is common to all engineering work. The problem may involve quantitative or qualitative factors; it may be physical or economic; it may require abstract mathematics or common sense. Of great importance is the process of creative synthesis or design, putting ideas together to create a new and optimum solution.

Although engineering problems vary in scope and complexity, the same general approach is applicable. First comes an analysis of the situation and a preliminary decision on a plan of attack. In line with thisplan, the problem is reduced to a more categorical question that can be clearly stated. The stated question is then answered by deductive reasoning from known principles or by creative synthesis, as in a new design. The answer or design is always checked for accuracy and adequacy. Finally, the results for the simplified problem are interpreted in terms of the original problem and reported in an appropriate form.

In order of decreasing emphasis on science, the major functions of all engineering branches are the following:

Research. Using mathematical and scientific concepts, experimental techniques, and inductive reasoning, the research engineer seeks new principles and processes.

Development. Development engineers apply the results of research to useful purposes. Creative application of new knowledge may result in a working model of a new electrical circuit, a chemical process, or an industrial machine.

Design. In designing a structure or a product, the engineer selects methods, specifies materials, and determines shapes to satisfy technical requirements and to meet performance specifications.

Construction. The construction engineer is responsible for preparing the site, determining procedures that will economically and safely yield the desired quality, directing the placement of materials, and organizing the personnel and equipment.

Production. Plant layout and equipment selection are the responsibility of the production engineer, who chooses processes and tools, integrates the flow of materials and components, and provides for testing and inspection.

Operation. The operating engineer controls machines, plants, and organizations providing power, transportation, and communication; determines procedures; and supervises personnel to obtain reliable and economic operation of complex equipment.

Management and other functions.

In some countries and industries, engineers analyze customers' requirements, recommend units to satisfy needs economically, and resolve related problems.

Major fields of engineering

MILITARY ENGINEERING

In its earliest uses the term engineering referred particularly to the construction of engines of war and the execution of works intended to serve military purposes. Military engineers were long the only ones to whom the title engineer was applied.

The role of the military engineer in modern war is to apply engineering knowledge and resources to the furtherance of the commander's plans. The basic requirement is a sound general engineering knowledge directed to the technical aspects of those tasks likely to be encountered in war. Engineering work is influenced by topographical considerations and in battle also by tactical limitations. At times engineering factors will actually govern the choice of the military plan adopted; a military engineer must, therefore, possess a sound military education so that the best technical advice will be given to the commander.

History.

In the prehistoric period every man was a fighter and every fighter was to some extent an engineer. Primitive efforts were restricted to the provision of artificial protection for the person and machines for hurling destruction at the enemy. In the earliest war annals it is difficult to distinguish the military from the civil engineer. Julius Caesar referred to his praefectus fabrum, an official who controlled the labour gangs employed on road making and also parties of artisans. The Domesday survey of AD 1086 included one "Waldivus Ingeniator," who held nine manors direct from the crown and was probably William the Conqueror's chief engineer in England. Throughout the Middle Ages, ecclesiastics were frequently employed as military engineers, not only for purposes of planning and building but also for fighting. One of the best known is Gundulph, bishop of Rochester, who built the White Tower of the Tower of London and Rochester Castle.

Thus, in ancient and medieval times the military engineer became a specialist who made and used engines of war such as catapults, ballistas, battering rams, ramps, towers, scaling ladders, and other devices in attacking or defending castles, fortresses, and fortified camps. In peacetime the military engineer built fortifications for the defense of the country or city. Because such engineers frequently dug trenches or tunnels as means of approaching or undermining enemy positions, they came to be called sappers or miners. With the invention of gunpowder and the countless other inventions that came in later centuries, the military engineer was required to have far more technical knowledge. He nevertheless remained a soldier and fought side by side with the infantry in many wars.

Before the late 17th century the engineers of French armies were selected infantry officers given brevets as engineers; they performed both civil and military duties for the king's service. In 1673 Sébastien Le Prestre de Vauban was appointed director general of the royal fortifications, and it was largely owing to this great designer of fortified places that in 1690 an officer corps of engineers was established. Sapper and miner companies were formed later, although these units

were generally attached to the artillery. In 1801 the officer corps of engineers was integrated with the sapper and miner units, and the amalgamated corps served with great distinction throughout Napoleon's campaigns. In 1868 military telegraphists were added to the corps. The first engineer railway battalion was formed in 1876, and a battalion of aeronauts raised in 1904 was the forerunner of the French air force.

The first military engineering school was established at Mézières in 1748, and Lazare Carnot, a former graduate of Mézières, moved the school to Metz in 1795, where it was renamed the École Polytechnique ("Polytechnic School").

Military engineering functions.

The functions of modern military engineers vary among the armies of the world, but as a rule they include the following activities: (1) construction and maintenance of roads, bridges, airfields, landing strips, and zones for the airdrop of personnel and supplies, (2) interference with the enemy's mobility by means of demolitions, floods, destruction of matériel, mine fields, and obstacles and fortifications of many types, (3) mapping and aiding the artillery to survey gun positions, rocket-launching sites, and target areas, (4) supplying water and engineering equipment, and (5) disposal of unexploded bombs or warheads. In the British army the Royal Engineers also operate the army postal service.

The U.S. Army Corps of Engineers is both a combatant arm and a technical service. Alone among the arms and services, it engages in civil as well as military activities. During the 20th century its civil works activities have centred upon the planning, construction, and maintenance of improvements to rivers, harbours, and other waterways and upon flood control. The principal military service performed by the Corps of Engineers in the United States and abroad is the construction and maintenance of buildings and utilities. In theatres of operation in wartime, such construction is carried out by engineer troops. In the United States in peace and war and overseas in peacetime, such construction is usually accomplished by private industry under contract to the Corps of Engineers.

CIVIL ENGINEERING

The term civil engineering was first used in the 18th century to distinguish the newly recognized profession from military engineering, until then preeminent. From earliest times, however, engineers have engaged in peaceful activities, and many of the civil engineering works of ancient and medieval times - such as the Roman public baths, roads, bridges, and aqueducts; the Flemish canals; the Dutch sea defenses; the French Gothic cathedrals; and many other monuments - reveal a history of inventive genius and persistent experimentation.

History. The beginnings of civil engineering as a separate discipline may be seen in the foundation in France in 1716 of the Bridge and Highway Corps, out of which in 1747 grew the École Nationale des Ponts et Chaussées ("National School of Bridges and Highways"). Its teachers wrote books that became standard works on the mechanics of materials, machines, and hydraulics, and leading British engineers learned French to read them. As design and calculation replaced rule of thumb and empirical formulas, and as expert knowledge was codified and formulated, the nonmilitary engineer moved to the front of the stage. Talented, if often self-taught, craftsmen, stonemasons, millwrights, toolmakers, and instrument makers became civil engineers. In Britain, James Brindley began as a millwright and became the foremost canal builder of the century; John Rennie was a millwright's apprentice who eventually built the new London Bridge; Thomas Telford, a stonemason, became Britain's leading road builder.

John Smeaton, the first man to call himself a civil engineer, began as an instrument maker. His design of Eddystone Lighthouse (1756-59), with its interlocking masonry, was based on a craftsman's experience. Smeaton's work was backed by thorough research, and his services were much in demand. In 1771 he founded the Society of Civil Engineers (now known as the Smeatonian Society). Its object was to bring together experienced engineers, entrepreneurs, and lawyers to promote the building of large public works, such as canals (and later railways), and to secure the parliamentary powers necessary to execute their schemes. Their meetings were held during parliamentary sessions; the society follows this custom to this day.

The École Polytechnique was founded in Paris in 1794, and the Bauakademie was started in Berlin in 1799, but no such schools existed in Great Britain for another two decades. It was this lack of opportunity for scientific study and for the exchange of experiences that led a group of young men in 1818 to found the Institution of Civil Engineers. The founders were keen to learn from one another and from their elders, and in 1820 they invited Thomas Telford, by then the dean of British civil engineers, to be their first president. There were similar developments elsewhere. By the mid-19th century there were civil engineering societies in many European countries and the United States, and the following century produced similar institutions in almost every country in the world.

Formal education in engineering science became widely available as other countries followed the lead of France and Germany. In Great Britain the universities, traditionally seats of classical learning, were reluctant to embrace the new disciplines. University College, London, founded in 1826, provided a broad range of academic studies and offered a course in mechanical philosophy. King's College, London, first taught civil engineering in 1838, and in 1840 Queen Victoria founded the first chair of civil engineering and mechanics at the University of Glasgow, Scot. Rensselaer Polytechnic Institute, founded in 1824, offered the first courses in civil engineering in the United States. The number of universities throughout the world with engineering faculties, including civil engineering, increased rapidly in the 19th and early 20th centuries. Civil engineering today is taught in universities on every continent.

Civil engineering functions. The functions of the civil engineer can be divided into three categories: those performed before construction (feasibility studies, site investigations, and design), those performed during construction (dealing with clients, consulting engineers, and contractors), and those performed after construction (maintenance and research).

Feasibility studies. No major project today is started without an extensive study of the objective and without preliminary studies of possible plans leading to a recommended scheme, perhaps with alternatives. Feasibility studies may cover

alternative methods - e.g., bridge versus tunnel, in the case of a water crossing - or, once the method is decided, the choice of route. Both economic and engineering problems must be considered.

Site investigations. A preliminary site investigation is part of the feasibility study, but once a plan has been adopted a more extensive investigation is usually imperative. Money spent in a rigorous study of ground and substructure may save large sums later in remedial works or in changes made necessary in constructional methods.

Since the load-bearing qualities and stability of the ground are such important factors in any large-scale construction, it is surprising that a serious study of soil mechanics did not develop until the mid-1930s. Karl von Terzaghi, the chief founder of the science, gives the date of its birth as 1936, when the First International Conference on Soil Mechanics and Foundation Engineering was held at Harvard University and an international society was formed. Today there are specialist societies and journals in many countries, and most universities that have a civil engineering faculty have courses in soil mechanics.

Design. The design of engineering works may require the application of design theory from many fields - e.g., hydraulics, thermodynamics, or nuclear physics. Research in structural analysis and the technology of materials has opened the way for more rational designs, new design concepts, and greater economy of materials. The theory of structures and the study of materials have advanced together as more and more refined stress analysis of structures and systematic testing has been done. Modern designers not only have advanced theories and readily available design data, but structural designs can now be rigorously analyzed by computers.

Construction. The promotion of civil engineering works may be initiated by a private client, but most work is undertaken for large corporations, government authorities, and public boards and authorities. Many of these have their own engineering staffs, but for large specialized projects it is usual to employ consulting engineers.

The consulting engineer may be required first to undertake feasibility studies, then to recommend a scheme and quote an approximate cost. The engineer is responsible for the design of the works, supplying specifications, drawings, and legal documents in sufficient detail to seek competitive tender prices. The engineer must compare quotations and recommend acceptance of one of them. Although he is not a party to the contract, the engineer's duties are defined in it; the staff must supervise the construction and the engineer must certify completion of the work. Actions must be consistent with duty to the client; the professional organizations exercise disciplinary control over professional conduct. The consulting engineer's senior representative on the site is the resident engineer.

A phenomenon of recent years has been the turnkey or package contract, in which the contractor undertakes to finance, design, specify, construct, and commission a project in its entirety. In this case, the consulting engineer is engaged by the contractor rather than by the client.

The contractor is usually an incorporated company, which secures the contract on the basis of the consulting engineer's specification and general drawings. The consulting engineer must agree to any variations introduced and must approve the detailed drawings.

Maintenance.

The contractor maintains the works to the satisfaction of the consulting engineer. Responsibility for maintenance extends to ancillary and temporary works where these form part of the overall construction.

After construction a period of maintenance is undertaken by the contractor, and the payment of the final installment of the contract price is held back until released by the consulting engineer. Central and local government engineering and public works departments are concerned primarily with maintenance, for which they employ direct labour.

Research. Research in the civil engineering field is undertaken by government agencies, industrial foundations, the universities, and other institutions. Most countries have government-controlled agencies, such as the United States Bureau

of Standards and the National Physical Laboratory of Great Britain, involved in a broad spectrum of research, and establishments in building research, roads and highways, hydraulic research, water pollution, and other areas. Many are government-aided but depend partly on income from research work promoted by industry.

Branches of civil engineering. In 1828 Thomas Tredgold of England wrote:

The most important object of Civil Engineering is to improve the means of production and of traffic in states, both for external and internal trade. It is applied in the construction and management of roads, bridges, railroads, aqueducts, canals, river navigation, docks and storehouses, for the convenience of internal intercourse and exchange; and in the construction of ports, harbours, moles, breakwaters and lighthouses; and in the navigation by artificial power for the purposes of commerce.

It is applied to the protection of property where natural powers are the sources of injury, as by embankments for the defence of tracts of country from the encroachments of the sea, or the overflowing of rivers; it also directs the means of applying streams and rivers to use, either as powers to work machines, or as supplies for the use of cities and towns, or for irrigation; as well as the means of removing noxious accumulations, as by the drainage of towns and districts to secure the public health.

A modern description would include the production and distribution of energy, the development of aircraft and airports, the construction of chemical process plants and nuclear power stations, and water desalination. These aspects of civil engineering may be considered under the following headings: construction, transportation, maritime and hydraulic engineering, power, and public health.

Construction. Almost all civil engineering contracts include some element of construction work. The development of steel and concrete as building materials had the effect of placing design more in the hands of the civil engineer than the architect. The engineer's analysis of a building problem, based on function and economics, determines the building's structural design.

Transportation. Roman roads and bridges were products of military engineering, but the pavements of McAdam and the bridges of Perronet were the work of the civil engineer. So were the canals of the 18th century and the railways of the 19th, which, by providing bulk transport with speed and economy, lent a powerful impetus to the Industrial Revolution. The civil engineer today is concerned with an even larger transportation field - e.g., traffic studies, design of systems for road, rail, and air, and construction including pavements, embankments, bridges, and tunnels.

Maritime and hydraulic engineering. Harbour construction and shipbuilding are ancient arts. For many developing countries today the establishment of a large, efficient harbour is an early imperative, to serve as the inlet for industrial plant and needed raw materials and the outlet for finished goods. In developed countries the expansion of world trade, the use of larger ships, and the increase in total tonnage call for more rapid and efficient handling. Deeper berths and alongside-handling equipment (for example, for ore) and navigation improvements are the responsibility of the civil engineer.

The development of water supplies was a feature of the earliest civilizations, and the demand for water continues to rise today. In developed countries the demand is for industrial and domestic consumption, but in many parts of the world - e.g., the Indus basin - vast schemes are under construction, mainly for irrigation to help satisfy the food demand, and are often combined with hydroelectric power generation to promote industrial development.

Dams today are among the largest construction works, and design development is promoted by bodies like the International Commission on Large Dams. The design of large impounding dams in places with population centres close by requires the utmost in safety engineering, with emphasis on soil mechanics and stress analysis. Most governments exercise statutory control of engineers qualified to design and inspect dams.

Power. Civil engineers have always played an important part in mining for coal and metals; the driving of tunnels is a task common to many branches of civil

engineering. In the 20th century the design and construction of power stations has advanced with the rapid rise in demand for electric power, and nuclear power stations have added a whole new field of design and construction, involving prestressed concrete pressure vessels for the reactor.

The exploitation of oil fields and the discoveries of natural gas in significant quantities have initiated a radical change in gas production. Shipment in liquid form from the Sahara and piping from the bed of the North Sea have been among the novel developments.

Public health. Drainage and liquid-waste disposal are closely associated with antipollution measures and the re-use of water. The urban development of parts of water catchment areas can alter the nature of runoff, and the training and regulation of rivers produce changes in the pattern of events, resulting in floods and the need for flood prevention and control.

Modern civilization has created problems of solid-waste disposal, from the manufacture of durable goods, such as automobiles and refrigerators, produced in large numbers with a limited life, to the small package, previously disposable, now often indestructible. The civil engineer plays an important role in the preservation of the environment, principally through design of works to enhance rather than to damage or pollute.

MECHANICAL ENGINEERING

Mechanical engineering is the branch of engineering that deals with machines and the production of power. It is particularly concerned with forces and motion.

History. The invention of the steam engine in the latter part of the 18th century, providing a key source of power for the Industrial Revolution, gave an enormous impetus to the development of machinery of all types. As a result, a new major classification of engineering dealing with tools and machines developed, receiving formal recognition in 1847 in the founding of the Institution of Mechanical Engineers in Birmingham, Eng.

Mechanical engineering has evolved from the practice by the mechanic of an art based largely on trial and error to the application by the professional engineer of

the scientific method in research, design, and production. The demand for increased efficiency is continually raising the quality of work expected from a mechanical engineer and requiring a higher degree of education and training.

Mechanical engineering functions.

Four functions of the mechanical engineer, common to all branches of mechanical engineering, can be cited. The first is the understanding of and dealing with the bases of mechanical science. These include dynamics, concerning the relation between forces and motion, such as in vibration; automatic control; thermodynamics, dealing with the relations among the various forms of heat, energy, and power; fluid flow; heat transfer; lubrication; and properties of materials.

Second is the sequence of research, design, and development. This function attempts to bring about the changes necessary to meet present and future needs. Such work requires a clear understanding of mechanical science, an ability to analyze a complex system into its basic factors, and the originality to synthesize and invent.

Third is production of products and power, which embraces planning, operation, and maintenance. The goal is to produce the maximum value with the minimum investment and cost while maintaining or enhancing longer term viability and reputation of the enterprise or the institution.

Fourth is the coordinating function of the mechanical engineer, including management, consulting, and, in some cases, marketing.

In these functions there is a long continuing trend toward the use of scientific instead of traditional or intuitive methods. Operations research, value engineering, and PABLA (problem analysis by logical approach) are typical titles of such rationalized approaches. Creativity, however, cannot be rationalized. The ability to take the important and unexpected step that opens up new solutions remains in mechanical engineering, as elsewhere, largely a personal and spontaneous characteristic.

Branches of mechanical engineering.

Development of machines for the production of goods.

The high standard of living in the developed countries owes much to mechanical engineering. The mechanical engineer invents machines to produce goods and develops machine tools of increasing accuracy and complexity to build the machines.

The principal lines of development of machinery have been an increase in the speed of operation to obtain high rates of production, improvement in accuracy to obtain quality and economy in the product, and minimization of operating costs. These three requirements have led to the evolution of complex control systems.

The most successful production machinery is that in which the mechanical design of the machine is closely integrated with the control system. A modern transfer (conveyor) line for the manufacture of automobile engines is a good example of the mechanization of a complex series of manufacturing processes. Developments are in hand to automate production machinery further, using computers to store and process the vast amount of data required for manufacturing a variety of components with a small number of versatile machine tools.

Development of machines for the production of power.

The steam engine provided the first practical means of generating power from heat to augment the old sources of power from muscle, wind, and water. One of the first challenges to the new profession of mechanical engineering was to increase thermal efficiencies and power; this was done principally by the development of the steam turbine and associated large steam boilers. The 20th century has witnessed a continued rapid growth in the power output of turbines for driving electric generators, together with a steady increase in thermal efficiency and reduction in capital cost per kilowatt of large power stations. Finally, mechanical engineers acquired the resource of nuclear energy, whose application has demanded an exceptional standard of reliability and safety involving the solution of entirely new problems (see Nuclear engineering below).

The mechanical engineer is also responsible for the much smaller internal combustion engines, both reciprocating (gasoline and diesel) and rotary (gas-

turbine and Wankel) engines, with their widespread transport applications. In the transportation field generally, in air and space as well as on land and sea, the mechanical engineer has created the equipment and the power plant, collaborating increasingly with the electrical engineer, especially in the development of suitable control systems.

Development of military weapons.

The skills applied to war by the mechanical engineer are similar to those required in civilian applications, though the purpose is to enhance destructive power rather than to raise creative efficiency. The demands of war have channeled huge resources into technical fields, however, and led to developments that have profound benefits in peace. Jet aircraft and nuclear reactors are notable examples.

Environmental control.

The earliest efforts of mechanical engineers were aimed at controlling the human environment by draining and irrigating land and by ventilating mines. Refrigeration and air conditioning are examples of the use of modern mechanical devices to control the environment.

Many of the products of mechanical engineering, together with technological developments in other fields, give rise to noise, the pollution of water and air, and the dereliction of land and scenery. The rate of production, both of goods and power, is rising so rapidly that regeneration by natural forces can no longer keep pace. A rapidly growing field for mechanical engineers and others is environmental control, comprising the development of machines and processes that will produce fewer pollutants and of new equipment and techniques that can reduce or remove the pollution already generated.

CHEMICAL ENGINEERING

Chemical engineering is the development of processes and the design and operation of plants in which materials undergo changes in physical or chemical state on a technical scale. Applied throughout the process industries, it is founded on the principles of chemistry, physics, and mathematics. The laws of physical chemistry and physics govern the practicability and efficiency of chemical

engineering operations. Energy changes, deriving from thermodynamic considerations, are particularly important. Mathematics is a basic tool in optimization and modeling. Optimization means arranging materials, facilities, and energy to yield as productive and economical an operation as possible. Modeling is the construction of theoretical mathematical prototypes of complex process systems, commonly with the aid of computers.

History.

Chemical engineering is as old as the process industries. Its heritage dates from the fermentation and evaporation processes operated by early civilizations. Modern chemical engineering emerged with the development of large-scale, chemical-manufacturing operations in the second half of the 19th century. Throughout its development as an independent discipline, chemical engineering has been directed toward solving problems of designing and operating large plants for continuous production.

Manufacture of chemicals in the mid-19th century consisted of modest craft operations. Increase in demand, public concern at the emission of noxious effluents, and competition between rival processes provided the incentives for greater efficiency. This led to the emergence of combines with resources for larger operations and caused the transition from a craft to a science-based industry. The result was a demand for chemists with knowledge of manufacturing processes, known as industrial chemists or chemical technologists. The term chemical engineer was in general use by about 1900. Despite its emergence in traditional chemicals manufacturing, it was through its role in the development of the petroleum industry that chemical engineering became firmly established as a unique discipline. The demand for plants capable of operating physical separation processes continuously at high levels of efficiency was a challenge that could not be met by the traditional chemist or mechanical engineer.

A landmark in the development of chemical engineering was the publication in 1901 of the first textbook on the subject, by George E. Davis, a British chemical consultant. This concentrated on the design of plant items for specific operations.

The notion of a processing plant encompassing a number of operations, such as mixing, evaporation, and filtration, and of these operations being essentially similar, whatever the product, led to the concept of unit operations. This was first enunciated by the American chemical engineer Arthur D. Little in 1915 and formed the basis for a classification of chemical engineering that dominated the subject for the next 40 years. The number of unit operations - the building blocks of a chemical plant - is not large. The complexity arises from the variety of conditions under which the unit operations are conducted.

In the same way that a complex plant can be divided into basic unit operations, so chemical reactions involved in the process industries can be classified into certain groups, or unit processes (e.g., polymerizations, esterifications, and nitrations), having common characteristics. This classification into unit processes brought rationalization to the study of process engineering.

The unit approach suffered from the disadvantage inherent in such classifications: a restricted outlook based on existing practice. Since World War II, closer examination of the fundamental phenomena involved in the various unit operations has shown these to depend on the basic laws of mass transfer, heat transfer, and fluid flow. This has given unity to the diverse unit operations and has led to the development of chemical engineering science in its own right; as a result, many applications have been found in fields outside the traditional chemical industry.

Study of the fundamental phenomena upon which chemical engineering is based has necessitated their description in mathematical form and has led to more sophisticated mathematical techniques. The advent of digital computers has allowed laborious design calculations to be performed rapidly, opening the way to accurate optimization of industrial processes. Variations due to different parameters, such as energy source used, plant layout, and environmental factors, can be predicted accurately and quickly so that the best combination can be chosen.

Chemical engineering functions.

Chemical engineers are employed in the design and development of both processes and plant items. In each case, data and predictions often have to be obtained or confirmed with pilot experiments. Plant operation and control is increasingly the sphere of the chemical engineer rather than the chemist. Chemical engineering provides an ideal background for the economic evaluation of new projects and, in the plant construction sector, for marketing.

Branches of chemical engineering.

The fundamental principles of chemical engineering underlie the operation of processes extending well beyond the boundaries of the chemical industry, and chemical engineers are employed in a range of operations outside traditional areas. Plastics, polymers, and synthetic fibres involve chemical-reaction engineering problems in their manufacture, with fluid flow and heat transfer considerations dominating their fabrication. The dyeing of a fibre is a mass-transfer problem. Pulp and paper manufacture involve considerations of fluid flow and heat transfer. While the scale and materials are different, these again are found in modern continuous production of foodstuffs. The pharmaceuticals industry presents chemical engineering problems, the solutions of which have been essential to the availability of modern drugs. The nuclear industry makes similar demands on the chemical engineer, particularly for fuel manufacture and reprocessing. Chemical engineers are involved in many sectors of the metals processing industry, which extends from steel manufacture to separation of rare metals.

Further applications of chemical engineering are found in the fuel industries. In the second half of the 20th century, considerable numbers of chemical engineers have been involved in space exploration, from the design of fuel cells to the manufacture of propellants. Looking to the future, it is probable that chemical engineering will provide the solution to at least two of the world's major problems: supply of adequate fresh water in all regions through desalination of seawater and environmental control through prevention of pollution.

ELECTRICAL AND ELECTRONICS ENGINEERING

Electrical engineering deals with the practical applications of electricity in all its forms, including those of the field of electronics. Electronics engineering is that branch of electrical engineering concerned with the uses of the electromagnetic spectrum and with the application of such electronic devices as integrated circuits, transistors, and vacuum tubes. In engineering practice, the distinction between electrical engineering and electronics is based on the comparative strength of the electric currents used. In this sense, electrical engineering is the branch dealing with "heavy current" - that is, electric light and power systems and apparatuses - whereas electronics engineering deals with such "light current" applications as wire and radio communication, the stored-program electronic computer, radar, and automatic control systems.

The distinction between the fields has become less sharp with recent technical progress. For example, in the high-voltage transmission of electric power, large arrays of electronic devices are used to convert transmission-line current at power levels in the tens of megawatts. Moreover, in the regulation and control of interconnected power systems, electronic computers are used to compute requirements much more rapidly and accurately than is possible by manual methods.

History.

Electrical phenomena attracted the attention of European thinkers as early as the 17th century. Beginning as a mathematically oriented science, the field has remained primarily in that form; mathematical predication often precedes laboratory demonstration. The most noteworthy pioneers include Ludwig Wilhelm Gilbert and Georg Simon Ohm of Germany, Hans Christian Ørsted of Denmark, André-Marie Ampère of France, Alessandro Volta of Italy, Joseph Henry of the United States, and Michael Faraday of England. Electrical engineering may be said to have emerged as a discipline in 1864 when the Scottish physicist James Clerk Maxwell summarized the basic laws of electricity in mathematical form and predicted that radiation of electromagnetic energy would occur in a form that later

became known as radio waves. In 1887 the German physicist Heinrich Hertz experimentally demonstrated the existence of radio waves.

The first practical application of electricity was the telegraph, invented by Samuel F.B. Morse in 1837. The need for electrical engineers was not felt until some 40 years later, upon the invention of the telephone (1876) by Alexander Graham Bell and of the incandescent lamp (1878) by Thomas A. Edison. These devices and Edison's first central generating plant in New York City (1882) created a large demand for men trained to work with electricity.

The discovery of the "Edison effect," a flow of current through the vacuum of one of his lamps, was the first observation of current in space. Hendrick Antoon Lorentz of The Netherlands predicted the electron theory of electrical charge in 1895, and in 1897 J.J. Thomson of England showed that the Edison effect current was indeed caused by negatively charged particles (electrons).

This led to the work of Guglielmo Marconi of Italy, Lee De Forest of the United States, and many others, which laid the foundations of radio engineering. In 1930 the term electronics was introduced to embrace radio and the industrial applications of electron tubes. Since 1947, when the transistor was invented by John Bardeen, William H. Brattain, and William B. Shockley, electronics engineering has been dominated by the applications of such solid-state electronic devices as the transistor, the semiconductor diode, and the integrated circuit.

Electrical and electronics engineering functions.

Research.

The functions performed by electrical and electronics engineers include (1) basic research in physics, other sciences, and applied mathematics in order to extend knowledge applicable to the field of electronics, (2) applied research based on the findings of basic research and directed at discovering new applications and principles of operation, (3) development of new materials, devices, assemblies, and systems suitable for existing or proposed product lines, (4) design of devices, equipment, and systems for manufacture, (5) field-testing of equipment and systems, (6) establishment of quality control standards to be observed in

manufacture, (7) supervision of manufacture and production testing, (8) postproduction assessment of performance, maintenance, and repair, and (9) engineering management, or the direction of research, development, engineering, manufacture, and marketing and sales.

Consulting. The rapid proliferation of new discoveries, products, and markets in the electrical and electronics industries has made it difficult for workers in the field to maintain the range of skills required to manage their activities. Consulting engineers, specializing in new fields, are employed to study and recommend courses of action.

The educational background required for these functions tends to be highest in basic and applied research. In most major laboratories a doctorate in science or engineering is required to fill leadership roles. Most positions in design, product development, and supervision of manufacture and quality control require a master's degree. In the high-technology industries typical of modern electronics, an engineering background at not less than the bachelor's level is required to assess competitive factors in sales engineering to guide marketing strategy.

Branches of electrical and electronics engineering.

The largest of the specialized branches of electrical engineering, the branch concerned with the electronic computer, was introduced during World War II. The field of computer science and engineering has attracted members of several disciplines outside electronics, notably logicians, linguists, and applied mathematicians.

Another very large field is that concerned with electric light and power and their applications. Specialities within the field include the design, manufacture, and use of turbines, generators, transmission lines, transformers, motors, lighting systems, and appliances.

A third major field is that of communications, which comprises not only telegraphy and telephony but also satellite communications and the transmission of voice and data by laser signals through optical-fibre networks. The communication of digital data among computers connected by wire, microwave, and satellite circuits is now

a major enterprise that has built a strong bond between computer and communications specialists.

The applications of electricity and electronics to other fields of science have expanded since World War II. Among the sciences represented are medicine, biology, oceanography, geoscience, nuclear science, laser physics, sonics and ultrasonics, and acoustics. Theoretical specialties within electronics include circuit theory, information theory, radio-wave propagation, and microwave theory.

Another important speciality concerns improvements in materials and components used in electrical and electronics engineering, such as conductive, magnetic, and insulating materials and the semiconductors used in solid-state devices. One of the most active areas is the development of new electronic devices, particularly the integrated circuits used in computers and other digital systems.

The development of electronic systems - equipment for consumers, such as radios, television sets, stereo equipment, video games, and home computers - occupies a large number of engineers. Another field is the application of computers and radio systems to automobiles, ships, and other vehicles. The field of aerospace electronic systems includes navigation aids for aircraft, automatic pilots, altimeters, and radar for traffic control, blind landing, and collision prevention. Many of these devices are also widely used in the marine services.

PETROLEUM ENGINEERING

Petroleum engineering is a specialized engineering discipline whose origins lie in both mining engineering and geology. The petroleum engineer, whose aim is to extract gaseous and liquid hydrocarbon products from the earth, is concerned with drilling, producing, processing, and transporting these products and handling all the related economic and regulatory considerations.

History.

The foundations of petroleum engineering were established during the 1890s in California. There geologists were employed to correlate oil-producing zones and water zones from well to well to prevent extraneous water from entering oil-producing zones. From this came the recognition of the potential for applying

technology to oil-field development. The American Institute of Mining and Metallurgical Engineers (AIME) established a Technical Committee on Petroleum in 1914. In 1957 the name of the AIME was changed to the American Institute of Mining, Metallurgical, and Petroleum Engineers.

Petroleum technology courses were introduced at the University of Pittsburgh, Pa., in 1910 and included courses in oil and gas law and industry practices; in 1915 the university granted the first degree in petroleum engineering. Also in 1910 the University of California at Berkeley offered its first courses in petroleum engineering and in 1915 established a four-year curriculum in petroleum engineering. After these pioneering efforts, professional programs spread throughout the United States and other countries.

From 1900 to 1920 petroleum engineering focused on drilling problems, such as establishing casing points for water shutoff, designing casing strings, and improving the mechanical operations in drilling and well pumping. In the 1920s petroleum engineers sought means to improve drilling practices and to improve well design by use of proper tubing sizes, chokes, and packers. They designed new forms of artificial lift, primarily rod pumping and gas lift, and studied the ways in which methods of production affected gas-oil ratios and rates of production. The technology of drilling fluids was advanced, and directional drilling became a common practice.

The economic crisis that resulted from abundant discoveries in about 1930, notably in the giant East Texas Field, caused petroleum engineering to focus on the entire oil-water-gas reservoir system rather than on the individual well. Studying the optimum spacing of wells in an entire field led to the concept of reservoir engineering. During this period the mechanics of drilling and production were not neglected. Drilling penetration rates increased approximately 100 percent from 1932 to 1937.

Petrophysics (determination of fluid and rock characteristics) was introduced late in the 1930s. By 1940 electric logging had developed to the state that estimates could be made of oil and water saturations in the reservoir rocks.

After World War II, petroleum engineers continued to refine the techniques of reservoir analysis and petrophysics. The outstanding event of the 1950s was development of the offshore oil industry and a whole new technology. At first little was known of such matters as wave heights and wave forces. The oceanographer and marine engineer thus joined with the petroleum engineer to initiate design standards. Shallow-water drilling barges evolved into mobile platforms, then into jack-up barges, and finally into semi-submersible and floating drilling ships.

Branches of petroleum engineering.

During the evolution of petroleum engineering, the areas of specialization developed: drilling engineering, production engineering, reservoir engineering, and petrophysical engineering. In each specialization engineers from other disciplines (mechanical, civil, electrical, geological, chemical) freely entered, and their contributions were significant; however, it remained the unique role of the petroleum engineer to integrate all the specializations into an efficient system of oil and gas drilling, production, and processing.

Drilling engineering was among the first applications of technology to oil-field practices. The drilling engineer is responsible for the design of the earth-penetration techniques, the selection of casing and safety equipment, and, often, the direction of the operations. These functions involve understanding the nature of the rocks to be penetrated, the stresses in these rocks, and the techniques available to drill into and control the underground reservoirs. Because modern drilling involves organizing a vast array of machinery and materials, investing huge funds, and acknowledging the safety and welfare of the general public, the engineer must develop the skills of supervision, management, and negotiation.

The production engineer's work begins upon completion of the well - directing the selection of producing intervals and making arrangements for various accessories, controls, and equipment. Later his work involves controlling and measuring the produced fluids (oil, gas, and water), designing and installing gathering and storage systems, and delivering the raw products (gas and oil) to pipeline companies and other transportation agents. He is also involved in such matters as corrosion

prevention, well performance, and formation treatments to stimulate production. As in all branches of petroleum engineering, the production engineer cannot view the in-hole or surface processing problems in isolation but must fit solutions into the complete reservoir, well, and surface system.

Reservoir engineers are concerned with the physics of oil and gas distribution and their flow through porous rocks - the various hydrodynamic, thermodynamic, gravitational, and other forces involved in the rock-fluid system. They are responsible for analyzing the rock-fluid system, establishing efficient well-drainage patterns, forecasting the performance of the oil or gas reservoir, and introducing methods for maximum efficient production.

To understand the reservoir rock-fluid system, the drilling, production, and reservoir engineers draw assistance from the petrophysical, or formation-evaluation, engineer, who provides tools and analytical techniques for determining rock and fluid characteristics. The petrophysical engineer measures the acoustic, radioactive, and electrical properties of the rock-fluid system and takes samples of the rocks and well fluids to determine porosity, permeability, and fluid content in the reservoir.

AEROSPACE ENGINEERING

Aerospace engineering is the study of the design, development, and operation of vehicles operating in the Earth's atmosphere or in outer space. In 1958 the first definition of aerospace engineering appeared, considering the Earth's atmosphere and the space above it as a single realm for development of flight vehicles. Today the more encompassing aerospace definition has commonly replaced the terms aeronautical engineering and astronautical engineering.

The design of a flight vehicle demands a knowledge of many engineering disciplines. It is rare that one person takes on the entire task; instead, most companies have design teams specialized in the sciences of aerodynamics, propulsion systems, structural design, materials, avionics, and stability and control systems. No single design can optimize all of these sciences, but rather there exist

compromised designs that incorporate the vehicle specifications, available technology, and economic feasibility.

History. Aeronautical engineering.

The roots of aeronautical engineering can be traced to the early days of mechanical engineering, to inventors' concepts, and to the initial studies of aerodynamics, a branch of theoretical physics. The earliest sketches of flight vehicles were drawn by Leonardo da Vinci, who suggested two ideas for sustentation. The first was an ornithopter, a flying machine using flapping wings to imitate the flight of birds. The second idea was an aerial screw, the predecessor of the helicopter. Manned flight was first achieved in 1783, in a hot-air balloon designed by the French brothers Joseph-Michel and Jacques-Étienne Montgolfier. Aerodynamics became a factor in balloon flight when a propulsion system was considered for forward movement. Benjamin Franklin was one of the first to propose such an idea, which led to the development of the dirigible. The power-driven balloon was invented by Henri Gifford, a Frenchman, in 1852. The invention of lighter-than-air vehicles occurred independently of the development of aircraft. The breakthrough in aircraft development came in 1799 when Sir George Cayley, an English baron, drew an airplane incorporating a fixed wing for lift, an empennage (consisting of horizontal and vertical tail surfaces for stability and control), and a separate propulsion system. Because engine development was virtually nonexistent, Cayley turned to gliders, building the first successful one in 1849. Gliding flights established a data base for aerodynamics and aircraft design. Otto Lilienthal, a German scientist, recorded more than 2,000 glides in a five-year period, beginning in 1891. Lilienthal's work was followed by the American aeronaut Octave Chanute, a friend of the American brothers Orville and Wilbur Wright, the fathers of modern manned flight.

Following the first sustained flight of a heavier-than-air vehicle in 1903, the Wright brothers refined their design, eventually selling airplanes to the U.S. Army. The first major impetus to aircraft development occurred during World War I, when aircraft were designed and constructed for specific military missions,

including fighter attack, bombing, and reconnaissance. The end of the war marked the decline of military high-technology aircraft and the rise of civil air transportation. Many advances in the civil sector were due to technologies gained in developing military and racing aircraft. A successful military design that found many civil applications was the U.S. Navy Curtiss NC-4 flying boat, powered by four 400-horsepower V-12 Liberty engines. It was the British, however, who paved the way in civil aviation in 1920 with a 12-passenger Handley-Page transport. Aviation boomed after Charles A. Lindbergh's solo flight across the Atlantic Ocean in 1927. Advances in metallurgy led to improved strength-to-weight ratios and, coupled with a monocoque design, enabled aircraft to fly farther and faster. Hugo Junkers, a German, built the first all-metal monoplane in 1910, but the design was not accepted until 1933, when the Boeing 247-D entered service. The twin-engine design of the latter established the foundation of modern air transport.

The advent of the turbine-powered airplane dramatically changed the air transportation industry. Germany and Britain were concurrently developing the jet engine, but it was a German Heinkel He 178 that made the first jet flight on Aug. 27, 1939. Even though World War II accelerated the growth of the airplane, the jet aircraft was not introduced into service until 1944, when the British Gloster Meteor became operational, shortly followed by the German Me 262. The first practical American jet was the Lockheed F-80, which entered service in 1945.

Commercial aircraft after World War II continued to use the more economical propeller method of propulsion. The efficiency of the jet engine was increased, and in 1949 the British de Havilland Comet inaugurated commercial jet transport flight. The Comet, however, experienced structural failures that curtailed the service, and it was not until 1958 that the highly successful Boeing 707 jet transport began nonstop transatlantic flights. While civil aircraft designs utilize most new technological advancements, the transport and general aviation configurations have changed only slightly since 1960. Because of escalating fuel and hardware prices, the development of civil aircraft has been dominated by the need for economical operation.

Technological improvements in propulsion, materials, avionics, and stability and controls have enabled aircraft to grow in size, carrying more cargo faster and over longer distances. While aircraft are becoming safer and more efficient, they are also now very complex. Today's commercial aircraft are among the most sophisticated engineering achievements of the day.

Smaller, more fuel-efficient airliners are being developed. The use of turbine engines in light general aviation and commuter aircraft is being explored, along with more efficient propulsion systems, such as the propfan concept. Using satellite communication signals, onboard microcomputers can provide more accurate vehicle navigation and collision-avoidance systems. Digital electronics coupled with servo mechanisms can increase efficiency by providing active stability augmentation of control systems. New composite materials providing greater weight reduction; inexpensive one-man, lightweight, noncertified aircraft, referred to as ultralights; and alternate fuels such as ethanol, methanol, synthetic fuel from shale deposits and coal, and liquid hydrogen are all being explored. Aircraft designed for vertical and short takeoff and landing, which can land on runways one-tenth the normal length, are being developed. Hybrid vehicles such as the Bell XV-15 tilt-rotor already combine the vertical and hover capabilities of the helicopter with the speed and efficiency of the airplane. Although environmental restrictions and high operating costs have limited the success of the supersonic civil transport, the appeal of reduced traveling time justifies the examination of a second generation of supersonic aircraft.

Aerospace engineering.

The use of rocket engines for aircraft propulsion opened a new realm of flight to the aeronautical engineer. Robert H. Goddard, an American, developed, built, and flew the first successful liquid-propellant rocket on March 16, 1926. Goddard proved that flight was possible at speeds greater than the speed of sound and that rockets can work in a vacuum. The major impetus in rocket development came in 1938 when the American James Hart Wyld designed, built, and tested the first U.S. regeneratively cooled liquid rocket engine. In 1947 Wyld's rocket engine powered

the first supersonic research aircraft, the Bell X-1, flown by the U.S. Air Force captain Charles E. Yeager. Supersonic flight offered the aeronautical engineer new challenges in propulsion, structures and materials, high-speed aeroelasticity, and transonic, supersonic, and hypersonic aerodynamics. The experience gained in the X-1 tests led to the development of the X-15 research rocket plane, which flew more than 700 flights over a 22-year period. The X-15 established an extensive database in transonic and supersonic flight (up to five times the speed of sound) and revealed vital information concerning the upper atmosphere.

The late 1950s and '60s marked a period of intense growth for astronautical engineering. In 1957 the U.S.S.R. orbited Sputnik I, the world's first artificial satellite, which triggered a space exploration race with the United States. In 1961 U.S. president John F. Kennedy recommended to Congress to undertake the challenge of "landing a man on the Moon and returning him safely to the Earth" by the end of the 1960s. This commitment was fulfilled on July 20, 1969, when astronauts Neil A. Armstrong and Edwin E. Aldrin, Jr., landed on the Moon.

The 1970s began the decline of the U.S. manned spaceflights. The exploration of the Moon was replaced by unmanned voyages to Jupiter, Saturn, and other planets. The exploitation of space was redirected from conquering distant planets to providing a better understanding of the human environment. Artificial satellites provide data pertaining to geographic formations, oceanic and atmospheric movements, and worldwide communications. The frequency of U.S. spaceflights in the 1960s and '70s led to the development of a reusable, low-orbital-altitude space shuttle. Known officially as the Space Transportation System, the shuttle has made numerous flights since its initial launch on April 12, 1981. It has been used for both military and commercial purposes (e.g., deployment of communications satellites).

Aerospace engineering functions.

In most countries, governments are the aerospace industry's largest customers, and most engineers work on the design of military vehicles. The largest demand for aerospace engineers comes from the transport and fighter aircraft, missile,

spacecraft, and general aviation industries. The typical aerospace engineer holds a bachelor's degree, but there are many engineers holding master's or doctorate degrees (or their equivalents) in various disciplines associated with aerospace-vehicle design, development, and testing.

The U.S. National Aeronautics and Space Administration (NASA) is a governmental organization that employs many engineers for research, development, testing, and procurement of military vehicles. Government agencies award and monitor industrial contracts ranging from engineering problem studies to design and fabrication of hardware. Universities receive limited funding, primarily for analytical research. Some of the larger institutions, however, are developing or expanding flight-research facilities and increasing faculty members in an effort to increase productivity in both research and testing.

The design of a flight vehicle is a complex and time-consuming procedure requiring the integration of many engineering technologies. Supporting teams are formed to provide expertise in these technologies, resulting in a completed design that is the best compromise of all the engineering disciplines. Usually the support teams are supervised by a project engineer or chief designer for technical guidance and by a program manager responsible for program budgets and schedules. Because of the ever-increasing requirement for advanced technology and the high cost and high risk associated with complex flight vehicles, many research and development programs are canceled before completion.

The design process can be dissected into five phases and is the same for most aerospace products. Phase one is a marketing analysis to determine customer specifications or requirements. Aerospace engineers are employed to examine technical, operational, or financial problems. The customer's requirements are established and then passed on to the conceptual design team for the second phase.

The conceptual design team generally consists of aerospace engineers, who make the first sketch attempt to determine the vehicle's size and configuration. Preliminary estimates of the vehicle's performance, weight, and propulsion systems are made. Performance parameters include range, speed, drag, power required,

payload, and takeoff and landing distances. Parametric trade studies are conducted to optimize the design, but configuration details usually change. This phase may take from a few months to years for major projects.

Phase three is the preliminary design phase. The optimized vehicle design from phase two is used as the starting point. Aerospace engineers perform computer analyses on the configuration; then wind-tunnel models are built and tested. Flight control engineers study dynamic stability and control problems. Propulsion groups supply data necessary for engine selection. Interactions between the engine inlet and vehicle frame are studied. Civil, mechanical, and aerospace engineers analyze the bending loads, stresses, and deflections on the wing, airframe, and other components. Material science engineers aid in selecting low-weight, high-strength materials and may conduct aeroelastic and fatigue tests. Weight engineers make detailed estimates of individual component weights. As certain parameters drive the vehicle design, the preliminary designers are often in close contact with both the conceptual designers and the marketing analysts. The time involved in the preliminary design phase depends on the complexity of the problem but usually takes from six to 24 months.

Phase four, the detailed design phase, involves construction of a prototype. Mechanical engineers, technicians, and draftsmen help lay out the drawings necessary to construct each component.

Full-scale mock-ups are built of cardboard, wood, or other inexpensive materials to aid in the subsystem layout. Subsystem components are built and bench-tested, and additional wind-tunnel testing is performed. This phase takes from one to three years.

The final phase concerns flight-testing the prototype. Engineers and test pilots work together to assure that the vehicle is safe and performs as expected. If the prototype is a commercial transport aircraft, the vehicle must meet the requirements specified by government organizations such as the Federal Aviation Administration in the United States and the Civil Aviation Authority in the United Kingdom.

Prototype testing is usually completed in one year but can take much longer because of unforeseen contingencies. The time required from the perception of a customer's needs to delivery of the product can be as long as 10 to 15 years depending on the complexity of the design, the political climate, and the availability of funding.

High-speed computers have now enabled complex aerospace engineering problems to be analyzed rapidly. More extensive computer programs, many written by aerospace engineers, are being formulated to aid the engineer in designing new configurations.

Branches of aerospace engineering.

The aerospace engineer is armed with an extensive background suitable for employment in most positions traditionally occupied by mechanical engineers as well as limited positions in the other various engineering disciplines. The transportation, construction, communication, and energy industries provide the most opportunities for non-aerospace applications.

Because land and sea vehicles are designed for optimum speed and efficiency, the aerospace engineer has become a prominent member of the design teams. Because up to half of the power required to propel a vehicle is due to the resistance of the air, the configuration design of low-drag automobiles, trains, and boats offers better speed and fuel economy. The presence of the aerospace engineer in the automobile industry is evident from the streamlined shapes of cars and trucks that evolved during the late 20th century, at a time when gasoline prices were escalating and the aerospace industry was in a lull. Airline companies employ engineers as performance analysts, crash investigators, and consultants. The Federal Aviation Administration makes use of the technical expertise of the aerospace engineer in various capacities.

The construction of large towers, buildings, and bridges requires predictions of aerodynamic forces and the creation of an optimum design to minimize these forces. The consideration of aerodynamic forces of flat surfaces such as the side of a building or superstructure is not new. In 1910 Alexandre-Gustave Eiffel achieved

remarkable experimental results measuring the wind resistance of a flat plate, using the Eiffel Tower as a test platform.

Many companies benefit not from the advanced hardware developments of aerospace technology but by the understanding and application of aerospace methodology. Companies engaged in satellite communications require an understanding of orbital mechanics, trajectories, acceleration forces, and aerodynamic heating and an overall knowledge of the spacecraft industry. Advanced aerodynamic design of airfoils and rotor systems is applied in an effort to improve the efficiency of propellers, windmills, and turbine engines. The impact of aerospace technology has trickled down to many companies engaged in the research and development of flight simulation, automatic controls, materials, dynamics, robotics, medicine, and other high-technology fields.

BIOENGINEERING

Bioengineering is the application of engineering knowledge to the fields of medicine and biology. The bioengineer must be well grounded in biology and have engineering knowledge that is broad, drawing upon electrical, chemical, mechanical, and other engineering disciplines. The bioengineer may work in any of a large range of areas. One of these is the provision of artificial means to assist defective body functions - such as hearing aids, artificial limbs, and supportive or substitute organs. In another direction, the bioengineer may use engineering methods to achieve biosynthesis of animal or plant products - such as for fermentation processes.

History.

Before World War II the field of bioengineering was essentially unknown, and little communication or interaction existed between the engineer and the life scientist. A few exceptions, however, should be noted. The agricultural engineer and the chemical engineer, involved in fermentation processes, have always been bioengineers in the broadest sense of the definition since they deal with biological systems and work with biologists. The civil engineer, specializing in sanitation, has applied biological principles in the work. Mechanical engineers have worked with

the medical profession for many years in the development of artificial limbs. Another area of mechanical engineering that falls in the field of bioengineering is the air-conditioning field. In the early 1920s engineers and physiologists were employed by the American Society of Heating and Ventilating Engineers to study the effects of temperature and humidity on humans and to provide design criteria for heating and air-conditioning systems.

Today there are many more examples of interaction between biology and engineering, particularly in the medical and life-support fields. In addition to an increased awareness of the need for communication between the engineer and the associate in the life sciences, there is an increasing recognition of the role the engineer can play in several of the biological fields, including human medicine, and, likewise, an awareness of the contributions biological science can make toward the solution of engineering problems.

Much of the increase in bioengineering activity can be credited to electrical engineers. In the 1950s bioengineering meetings were dominated by sessions devoted to medical electronics. Medical instrumentation and medical electronics continue to be major areas of interest, but biological modeling, blood-flow dynamics, prosthetics, biomechanics (dynamics of body motion and strength of materials), biological heat transfer, biomaterials, and other areas are now included in conference programs.

Bioengineering developed out of specific desires or needs: the desire of surgeons to bypass the heart, the need for replacement organs, the requirement for life support in space, and many more. In most cases the early interaction and education were a result of personal contacts between physician, or physiologist, and engineer. Communication between the engineer and the life scientist was immediately recognized as a problem. Most engineers who wandered into the field in its early days probably had an exposure to biology through a high-school course and no further work. To overcome this problem, engineers began to study not only the subject matter but also the methods and techniques of their counterparts in medicine, physiology, psychology, and biology. Much of the information was self-

taught or obtained through personal association and discussions. Finally, recognizing a need to assist in overcoming the communication barrier as well as to prepare engineers for the future, engineering schools developed courses and curricula in bioengineering.

Branches of bioengineering.

Medical engineering. Medical engineering concerns the application of engineering principles to medical problems, including the replacement of damaged organs, instrumentation, and the systems of health care, including diagnostic applications of computers.

Agricultural engineering. This includes the application of engineering principles to the problems of biological production and to the external operations and environment that influence this production.

Bionics. Bionics is the study of living systems so that the knowledge gained can be applied to the design of physical systems.

Biochemical engineering. Biochemical engineering includes fermentation engineering, application of engineering principles to microscopic biological systems that are used to create new products by synthesis, including the production of protein from suitable raw materials.

Human-factors engineering.

This concerns the application of engineering, physiology, and psychology to the optimization of the human-machine relationship.

Environmental health engineering. Also called bioenvironmental engineering, this field concerns the application of engineering principles to the control of the environment for the health, comfort, and safety of human beings. It includes the field of life-support systems for the exploration of outer space and the ocean.

NUCLEAR ENGINEERING

Nuclear engineering is concerned with the control and use of energy and radiation released from nuclear reactions. It encompasses the development, design, and construction of power reactors, naval-propulsion reactors, nuclear fuel-cycle

facilities, and radioactive-waste disposal facilities; the development and production of nuclear weapons; and the production and application of radioisotopes.

History. Nuclear engineering began with the first major demonstrations of the utilization of nuclear energy: the development of nuclear weapons and nuclear reactors.

The World War II Manhattan Project, under which the U.S. government built, in a relatively short period, such facilities as production reactors, chemical-reprocessing plants, test and research reactors, and weapons production facilities, stands out as a monumental engineering feat. Engineers in early programs had to learn about a host of nuclear-related subjects, ranging from reactor theory and reactor control to radioactivity and the behaviour of material under irradiation. They were educated on the job by nuclear scientists and physicists, first through personal discussions and later through seminars and classes. Many of those who entered the field had been educated in other engineering disciplines - mechanical, electrical, chemical, and so on. Nuclear engineering continues today to be a strongly interdisciplinary activity.

Early schools. In the late 1940s, as the many potential peaceful uses of nuclear energy became evident, two schools of reactor technology were established, one in Tennessee at Oak Ridge National Laboratory and another in Illinois at Argonne National Laboratory.

In 1946 Clinch College was established at Oak Ridge. In its first year 35 American participants from universities, industry, the U.S. Navy, and government agencies took courses in nuclear technology. They attended lectures, conducted laboratory experiments, and gained hands-on experience in operating nuclear reactors.

In 1950 Clinch College was succeeded by the Oak Ridge School of Reactor Technology (ORSORT). The participants were again selected from academic, government, and industry sectors. In addition to lectures and laboratory work, the students were assigned to teams working on the development of new concepts. Several concepts developed by these teams later grew into major research and development programs, including the high-flux isotope reactor, the molten-salt

reactor, and several nuclear propulsion schemes. ORSORT was disbanded in 1965 because nuclear engineering programs had by that time become widely available at universities and colleges.

The International School of Nuclear Science and Engineering was established at Argonne National Laboratory in 1955. The school was created to meet the international need for trained scientists and engineers, and its program was conducted jointly by Argonne National Laboratory, North Carolina State College, and Pennsylvania State University. Basic course work was presented at the universities in a 17-week program combining lecture with laboratory experience. More advanced work, including lectures and participation in design and laboratory projects, was given in a second 17-week program at the International School at Argonne. In 1960 the basic course work was discontinued, and the program was redirected to serve more advanced and experienced students from abroad. In recognition of the worldwide growth of programs and facilities to provide basic nuclear training at universities and laboratories, the program at Argonne was discontinued in 1964.

University programs. In 1950 the first full-fledged nuclear engineering curriculum offered for college credit was established at North Carolina State College. By 1952 several schools had graduate programs in nuclear engineering.

Most of these programs consisted of two or three courses, providing a background on reactor physics, reactor control, heat transfer, radiation effects, and shielding.

With the support of the U.S. Atomic Energy Commission's Division of Nuclear Education and Training, the curricula and the number of schools in the United States continued to increase. By 1965, 61 schools were offering nuclear engineering programs. The programs had grown in diverse directions, however, and it became apparent that it was desirable to develop a consensus among educators about nuclear engineering education. To meet this need, a joint committee of the American Nuclear Society and the American Society of Engineering Education developed basic educational criteria. The committee members came from industry, national laboratories, and universities with nuclear

engineering programs. The committee's "Report on Objective Criteria in Nuclear Engineering Education" had a major influence in shaping nuclear engineering curricula around the world and did much to establish nuclear engineering as a distinct discipline.

Nuclear engineering functions.

Research and development. Research and development entails the conception and development of new materials, processes, components, and systems for nuclear facilities and the development of analytical methods and experimental procedures for use in the development, analysis, design, and control of fission and fusion systems.

Design. Another area of emphasis is the engineering design of such items as fuel elements, reactor-core supports, reflectors, thermal shields, biological shields, instrumentation and control systems, and safety systems.

Fuel management. Fuel management involves specifying, procuring, and managing fuel throughout its reactor lifetime and beyond.

Safety analysis. Normal and anticipated abnormal operating conditions must be considered in the analysis of the safety of a reactor or other facility using radioactive material. Hypothetical reactor accidents are analyzed to assess possible consequences and to devise means to prevent or mitigate these consequences.

Operation and test. This function of nuclear engineering is concerned with the supervision and operation of nuclear power reactors and ancillary nuclear facilities.

Nuclear engineers perform these functions for various kinds of employers: (1) architectural engineering firms, in which they handle design, safety analysis, project coordination, construction supervision, quality assurance, quality control, and related matters, (2) reactor vendors and other manufacturing organizations, in which they pursue research, development, design, manufacture, and installation of various components of nuclear systems, (3) electric utility companies, in which they handle planning, construction supervision, reactor-safety analysis, in-core nuclear fuel management, power-reactor economic analysis, environmental-impact assessment, personnel training, plant management, operation-shift supervision,

radiation protection, spent-fuel storage, and radioactive-waste management, (4) regulatory agencies, in which they undertake licensing, rule making, safety research, risk analysis, on-site inspection, and research administration, (5) defense programs, in which they are employed in naval and nuclear weapons programs, (6) universities, in which they hold various faculty positions, and (7) national laboratories and industrial research laboratories, in which they carry out advanced research and development on a variety of nuclear programs in nuclear energy areas. Most of the advanced research and development on nuclear-related programs is conducted at national laboratories.

Branches of nuclear engineering.

Nuclear power. The greatest growth in the nuclear industry has been in the development of nuclear power plants. It is estimated that by the year 2000 one-third of all electric power generated worldwide will come from nuclear power plants.

Nearly all commercial nuclear reactors in operation or under construction are thermal reactors. They are called thermal reactors because their fuel is fissioned by neutrons that have been slowed down by a moderator until they are in thermal equilibrium with the moderator. The boiling water reactor (BWR) and the pressurized water reactor (PWR) are the two predominant types of power reactors in use throughout the world. Both types are called light-water reactors (LWR). The water is used in these reactors as both moderator and coolant. In the BWR, steam is generated by direct boiling of water in the reactor core. In the PWR, steam is produced in an external steam generator rather than in the core, where the coolant under pressure is not allowed to boil. Other types of power reactors include graphite-moderated gas-cooled reactors in use in Great Britain and pressurized heavy-water reactors in Canada.

A major advance in nuclear power is expected with the further development of the liquid-metal fast-breeder reactor (LMFBR). Programs are in progress in several countries to develop and deploy the LMFBR. (The reactor is cooled by a liquid metal, sodium, and fission is caused by fast neutrons. The reactor is called a

breeder because it produces more nuclear fuel than it consumes.) Fuel in the breeder is utilized 60 times more effectively than that in light-water reactors. It is estimated that without the breeder the world supply of fissionable material for nuclear power plants could be consumed in a few decades. With the improved fuel utilization provided by the breeder, nuclear power plants would be able to supply the world's electric energy requirement for centuries.

Fusion. Fusion is a potential energy resource with a wide range of applications. The fusion process of combining two light atoms to form a heavier atom, with less mass than the two original atoms, is the basic energy process in the universe (i.e., fusion is the process that takes place in all stars). If fusion can be harnessed for terrestrial applications, the energy can be released in a variety of forms, including charged particles, electromagnetic radiation, and neutrons. Possible applications include electricity production, synthetic fuel production, process-heat applications, and fissile fuel production for fission reactors.

Fusion research since about 1950 has concentrated on the issues of plasma physics, specifically the production of high-temperature plasmas (100,000,000 C [180,000,000 F] or greater) that can be confined at sufficiently high densities for sufficiently long times to produce net energy. Energy break-even conditions are expected to be demonstrated in several fusion devices in the late 20th century. Fusion physics research has made steady progress, and research efforts have begun to address the important engineering issues of fusion. Among the more important of these issues are those related to extracting useful energy from a plasma and developing complete fuel systems for fusion reactors. These areas are expected to receive increased research and development support in the future.

Naval nuclear propulsion. The use of nuclear reactors to propel naval vessels has revolutionized naval operations throughout the world. The navies of Great Britain, France, China, the United States, Russia, and Ukraine are equipped with nuclear-powered ships, which are considered to be essential to the defense of their countries. Nuclear warships are capable of nearly unlimited high-speed operation without the need of fuel-oil support. In the 25 years following the maiden voyage

of the Nautilus in 1954, the nuclear navy of the United States steamed more than 80,000,000 kilometres (50,000,000 miles) throughout the oceans of the world, accumulating 25 centuries of reactor-plant operation without any accidents involving a nuclear reactor. By the mid-1980s, more than 40 percent of U.S. combat warships were nuclear-powered.

Nuclear weapons. Fission weapons (atomic bombs), fusion weapons (hydrogen bombs), and combination fission-fusion weapons are part of the world's nuclear arsenal. Nuclear engineers are employed on weapons programs in such diverse activities as research, development, design, fabrication, production, testing, maintenance, and surveillance of a large array of nuclear weapons systems.

Efforts are in progress in the United States to develop, upgrade, and integrate weapons into warhead programs and to explore advanced concepts for future weapons systems. A concept of particular interest is inertial-confinement fusion. This program is directed at determining the feasibility of burning very small pellets of thermonuclear fuel using laser or particle-beam drivers. The program is of interest not only for applications to weapons physics but also for possible energy applications.

Radioisotopes. More than 500 radioisotopes are produced in nuclear reactors. The production, packaging, and application of these isotopes has become a large industry. They are used in heart pacemakers, medical research, sterilization of medical instruments, industrial tracers, X-ray equipment, curing of plastics, preservation of food, and as an energy source in electric generators. Perhaps the most important use of radioisotopes is in the field of medicine. They are used in procedures for half of all patients admitted to hospitals in the United States.

Nuclear-waste management. Nuclear wastes can be classified in two groups, low-level and high-level. Low-level wastes come from nuclear power facilities, hospitals, and research institutions and include such items as contaminated clothing, wiping rags, tools, test tubes, needles, and other medical research materials. In the disposal of low-level wastes, the wastes are reduced in volume, then packaged in leak-proof containers, which are placed in an earth-covered

trench in a low-level-waste disposal site. Such sites should be continuously monitored to detect any migration of radioactive material. High-level wastes are highly radioactive and derive from the chemical reprocessing of spent fuel elements and from the weapons program.

By the late 20th century many countries were evaluating potential nuclear-waste disposal sites and developing terminal waste-storage technology. All these countries were preparing to handle high-level wastes. All had identified geologic formations that appeared to be technically feasible for repositories. In 1982 the U.S. Congress passed legislation establishing schedules for the selection, development, licensing, and construction of repositories for the safe, permanent storage of high-level waste.

ФАКУЛЬТЕТ ГУМАНИТАРНЫХ НАУК И СОЦИАЛЬНЫХ ТЕХНОЛОГИЙ

The Study of History

Modern historians aim to reconstruct a record of human activities and to achieve a more profound understanding of them. This conception of their task is quite recent, dating from the development in the late 18th and early 19th centuries of scientific history, cultivated largely by professional historians. It springs from an outlook that is very new in human experience: the assumption that the study of history is a natural, inevitable human activity. Before the late 18th century, historiography (the writing of history) did not stand at the centre of any civilization. History was almost never an important part of regular education, and it never claimed to provide an interpretation of human life as a whole. This was more appropriately the function of religion, of philosophy, even perhaps of poetry and other imaginative literature.

ANCIENT HISTORIOGRAPHY

Greco-Roman era.

The older, pre-18th-century outlook has been particularly well studied in the historiography of the ancient Greeks and Romans. But, although two of the most important ancient historians, Herodotus and Thucydides, wrote as early as the 5th century BC, when recorded Greek historiography was only just beginning, they had few successors of comparable quality. It is a symptom of the relative lack of importance attached in antiquity to this type of activity.

Ancient history was a branch of literature. The most appreciated historians were the writers who, like Thucydides, were able to touch onuniversal human problems or who, like the Roman author Tacitus (died c. AD 120), wrote in a dramatic way about important events or who, at least, attracted readers by their excellent style and skill in composition. Many of the works that lacked some of these literary qualities failed to survive.

About 1,000 ancient Greeks wrote in antiquity on historical subjects, but most of these writers are mere names. Many of the losses appear to have occurred in antiquity itself. Even historians of first rank have fared badly. Only in a few cases have complete texts of all their writings survived. Of the voluminous history of Polybius (covering originally the period 220-144 BC) only about one-third survives. Nearly half of Livy's Roman history (originally covering the period 753-9 BC) is lost. The text that remains is reasonably good only through the efforts of a group of Roman aristocrats who, in about AD 500, were trying to salvage the chief glories of Roman literature. A considerable part of Tacitus is missing, and the surviving portions of his Annals and Histories (originally AD 14-96) derive from two unique manuscripts.

Herodotus, whom the Roman statesman Cicero called "the father of history," came from the western coast of Asia Minor. The writers who preceded him were mainly Ionians from the Greek settlements in the same area. The origin of Greek historiography lies in the Ionian thought of the 6th century. The Ionian philosophers were doing something unprecedented: they were assuming that the universe is an intelligible whole and that through rational inquiries men might discover the general principles that govern it. Hecateus of Miletus, the most important Ionian predecessor of Herodotus, was applying the same critical spirit to the largely mythical Greek traditions when he wrote, early in the 5th century, "the stories of the Greeks are numerous and in my opinion ridiculous." Herodotus was more of a traditionalist, but he introduced his work as an "inquiry" (historia).

Egyptian and Babylonian historiography.

A glance at the older historiography of the Egyptians, the Babylonians, and the other peoples of the ancient Near East will heighten one's appreciation of the novelty of the task undertaken by Herodotus. The kings of Egypt, of Babylonia and Assyria, and of the Hittites and the Persians all sought to preserve their glorious deeds for posterity in monumental inscriptions. The more important rulers also accumulated large archives, including both ordinary administrative documents and records specially commemorating their achievements. Some 20,000 clay tablets

remain from the collections written for Ashurbanipal of Assyria (668-627 BC). Both in Egypt and in Babylonia lists of kings were kept in the temples, and these were sometimes supplemented by brief annals recording the principal events, though the hatred felt by certain rulers for their predecessors led to periodic destructions of older material. The exceptional meagreness of the narrative sources for Babylonian history before 747 BC seems due to the obliteration of the older annals by Nabonassar of Babylonia (ruled 747-734). Apart from changes in literary style, there was surprisingly little development over a period of more than 1,000 years in all these types of commemorative records. The inscriptions and temple records were normally intended to perpetuate the glory of the gods in whose service these rulers had accomplished great deeds. The names and dates of dynasties and of particular rulers can be reconstructed fairly adequately with the aid of these sources, but one cannot expect much accurate information about particular events. Nor, with rare exceptions, were those who had access to this material interested in using it to write continuous histories.

Herodotus and his immediate Ionian predecessors shared a very novel outlook. Its distinctive features were a lively curiosity and a capacity to treat sources in a critical spirit. Boundless curiosity about people and their diverse customs is one of the most endearing traits of Herodotus. Like other Greeks from western Asia Minor, he was particularly stimulated by contacts with the great Persian Empire, which offered opportunities for reasonably secure travel. The resultant immense widening of historical perspective is illustrated by a story told by Herodotus about Hecateus. When the latter assured the Egyptian priests at Thebes that he could trace his descent through 16 generations, the Egyptians showed him evidence of the descent of their high priests through 345 generations. Herodotus was the first to link his geographic inquiries with true history. His descriptions of the barbarian world that confronted the Greeks provided an introduction to the epic of the successful Greek resistance to the Persians.

Ancient history and biography.

The types of history written by the ancient Greeks and Romans influenced profoundly all subsequent historiography down to the 18th century. In order to interpret sympathetically this classical historiography, it is necessary to bear in mind the literary conventions that governed this branch of literature.

The ancient Greeks distinguished between history and biography. The origin of both forms can be traced back to at least the 5th century BC, and the differences between them were observed throughout antiquity. The writer of history was supposed to aim at giving a true story, but the biographer was entitled to treat historical personages in a manner that resembled legend. There existed, of course, some exceptions. The lives of the early Roman emperors written by Suetonius in the 2nd century AD, while conforming to the traditional, topical arrangement of biographies, constitute an unusually valuable historical source, especially for Augustus, whose correspondence is repeatedly quoted. Yet another distinction was drawn between history and the study of "antiquities," to use a term employed by Varro (116-27 BC), perhaps the greatest of all the ancient Roman scholars. This distinction was already implicit in Aristotle's contemptuous dismissal of history (in his Poetics) as a branch of literature dealing with the particular rather than with things of general significance. The histories he condemned provided chronological narratives of wars and political events. Aristotle and his disciples were engaged in several enterprises that they regarded as something quite different from history. For example, they embarked on the study of the constitutions of all the Greek states. Such work was to be based on systematic inquiries. The student of the "antiquities" tried to use a wider range of evidence than the sources normally consulted by the ancient historians, and he arranged his results systematically by topics.

In antiquity a writer of history was usually preoccupied at least as much with style as with content. A generation before Aristotle, the rules of rhetoric, as they might be applied to history, were fully elaborated by Isocrates, a teacher of rhetoric at Athens. Cicero tried (especially in his De oratore, 55 BC) to familiarize the Romans with these Isocratean precepts. History was to be written in a clear but

solemn style, akin to fine oratory. The historian was to introduce all manner of literary embellishments but was also to stress the moral lessons of his story. At its worst this type of historiography could lead to serious misrepresentations of the past. Among the Roman historians, Livy (died AD 17) was an important practitioner of this kind of writing, which was particularly well suited to the patriotic myths that he was trying to immortalize, of a Rome that owed its magnificent destiny to the unique virtues of its citizens and the perfection of its antique institutions. Some outstanding historians, such as Polybius (2nd century BC) and Caesar (died 44 BC), eschewed these rhetorical precepts, but in all the ancient writers an important element of literary artifice was always present. This is one of the reasons why they offend modern standards, which demand absolute accuracy in the presentation of evidence. One of the most striking contrasts is the reluctance of the ancient historians to quote documents. Tacitus might rely heavily on the archives of the Roman Senate, but he never mentions his documentary sources. An inscription discovered at Lyons, France, preserves a speech delivered by the emperor Claudius to the Senate in AD 48, and it is clear that Tacitus utilized another version of the same text. His skill in using it is matched by the freedom with which he adapts it to suit his purpose.

Methods of Thucydides.

The greatest and the most original achievement of the best Greek historians lay in their clear grasp of the need to distinguish truth from fiction and their conscious preoccupation with the methods of achieving this. This is admirably conveyed in a famous passage of Thucydides. And with reference to the narrative of events, far from permitting myself to derive it from the first source that came to hand, I did not even trust my own impressions, but it rests partly on what I saw myself, partly on what others saw for me, the accuracy of the report being always tried by the most severe and detailed tests possible. My conclusions have cost me some labour from the want of coincidence between accounts of the same occurrences by different eye-witnesses, arising sometimes from deficient memory, sometimes from deficient impartiality.

His practice did not fully live up to this ideal, however. The greatest of his Greek successors, Polybius, is reasonably impartial, except in his treatment of some of the events in Greece. Among the Romans, the writing of history was chiefly the preserve of members of the senatorial class, who almost invariably had some personal axes to grind. But the correctness of the rules formulated by Thucydides was accepted, in principle, by most ancient historians.

Thucydides had deliberately restricted himself to the history of his own time, and many of the subsequent ancient historians did likewise. They could depend on their own experience or could question well-informed contemporaries. The surviving fragments of Livy relating to his own lifetime (64/59 BC-AD 17) are much more vivid and convincing than the earlier books of his history (surviving today only down to 167 BC). The tendency to prefer contemporary history was strengthened by the practical bent of many of these writers. Several ancient historians were men of action familiar with warfare and politics. Interested in history as a source of instruction for statesmen, they could write with authority only about wars and political transactions of their own time. Polybius, the exiled Achaean general and a great traveller, derides unpractical, sedentary historians such as Timaeus, who had been writing about the peoples of the western Mediterranean without stirring for 50 years from Athens.

The historians of antiquity were much less skillful in dealing with noncontemporary history, for which they relied on older historians. Where none was to be found, they felt lost, as Livy complains in the early portions of his Roman history. The modern recourse to non-narrative sources was alien to the habits of most ancient historians. They were usually incapable of doing this successfully, just as they were ill equipped to discuss critically the sources used by the older writers.

Herodotus chose for his theme the successful resistance of the Greeks against the Persians at the beginning of the 5th century BC. Thucydides wrote about the Peloponnesian War, in which virtually all the Greek states became involved in the last decades of that century. These were limited subjects of obvious importance for

which it was possible to find ample evidence. The strength of the ancient historians lay precisely in imposing an interesting pattern on the events of a selected period, usually contemporary or fairly recent, for which they had manageable sources. The best of them could thereby achieve a sense of dramatic unity and produce literary masterpieces. The speeches that Thucydides invented for some of the main protagonists in his story are artistically the most satisfying parts of his work, and at times they even seem to recapture the spirit of what might have been said on these occasions. In a superb writer like Tacitus, whose political career had included long periods of frustration and insecurity, one does not look for impartiality or for scrupulous truthfulness but, rather, for fascinating insights into what the development of Roman imperial power from Augustus to Domitian (the period AD 14-96) meant to the proud, sophisticated Roman aristocracy for whom he was writing.

Classical study of "antiquities."

The study of "antiquities," as opposed to narrative history, did not normally produce works of literary merit, and this is probably the main reason why most of them disappeared. One important group of such writings originated with Aristotle and his collaborators, writing in the third quarter of the 4th century BC. They were interested in both literary "antiquities" and in the systematic study of the constitutions of Greek states. They had described 158 different constitutions, though only their account of Athens now survives. A comparison of its two main parts illustrates the contrast between the deficiencies of ancient historiography and the impressive achievements of the antiquarian researchers. In the introductory, historical section, Aristotle was baffled by the problem of dealing with the fairly remote past. For each particular period he tried to follow some contemporary sources. The resultant juxtaposition of several writers differing widely in their political outlook produced an account full of contradictions. The second part, however, containing a systematic description of the Athenian constitution, is a masterpiece of shrewd analysis, as are the empirical portions of Aristotle's Politics

(Books IV-VI), which are based on a wealth of concrete examples derived from the different Greek states.

Aristotle inspired in the 3rd and 2nd centuries BC a great mass of philological and antiquarian research. The most important scholars were to be found in the new Hellenistic states, especially at Alexandria in Egypt and at Pergamum in Asia Minor. Among the surviving Hellenistic fragments, there are commentaries on Herodotus and Thucydides. The Hellenistic scholars were interested in many subjects connected with history and did pioneering work in chronology, geography, and topography. They were accustomed to using every kind of source and to quoting documents extensively. Their greatest Roman disciple was Varro, who tried to recover all the vestiges of the old Roman society and to make a systematic survey of Roman life based on the evidence provided by language, literature, religion, and ancient customs. Most of his writings have been lost, but he supplied the conjectural (though incorrect) date of 753 BC for the foundation of Rome and knowledge of the probable boundaries between some of the groups whose union produced the city of Rome. Unfortunately, antiquarian researches of such penetrating nature were almost never applied in antiquity to the writing of narrative histories.

Early Christian era.

The triumph of Christianity in the Roman Empire during the 4th century assured the predominance of a type of historiography radically different from the works of the pagan Greek and Roman historians. Its origins were Jewish. The Jews were the only people of antiquity who had the supreme religious duty of remembering the past because their traditional histories commemorated the working out of God's plan for his chosen people. By contrast, no Greek ever heard his gods ordering him to remember. It was the duty of every Jew to be familiar with the Jewish sacred writings, which were ultimately gathered into what became the Old Testament. The writers of these biblical books only gave an authoritative version of what everybody was supposed to know, and they were only concerned with the selection of such facts as seemed relevant in interpreting God's purpose. In addition, the

Jews also cherished unwritten traditions. To quote Josephus, a Jewish historian of the 1st century AD, "what had not been written down, was yet entrusted to the collective memory of the people of Israel and especially of its priests."

The Christians took over the Old Testament and added to it an additional body of sacred history. he writers of the four Gospels included in the New Testament were bearing witness to assured truths that the faithful ought to know, and no convincing reconstruction of historical facts is possible from these books of the New Testament. The only avowedly historical book in it is the Acts of the Apostles. The New Testament as a whole represents merely a selection from the early Christian writings. It includes only what conformed to the doctrine of the church when, later on, that doctrine became fixed in one form. Between the Acts of the Apostles, dating probably from the late 1st century, and the writings of Eusebius of Caesarea (died c. 340) and his contemporaries in the first quarter of the 4th century, there is an almost complete gap in Christian historiography.

For the Christian writers the story of Jesus, as recorded in the Gospels, represented the fulfillment of the prophecies that could be found in various parts of the Old Testament. The Jewish part of the Bible also assured for Christianity the authority of a long antiquity. The history contained in the two parts of the Bible, now indissolubly linked together, became the only authentic record of God's revelation for mankind, dwarfing into insignificance all the records of other peoples and religious groups. The concept of a universal history had not been wholly unknown to the pagan world, but the Christians were the first to apply it effectively. Christian history had to be a universal history, though of a very peculiar sort, where only one sequence of privileged events, Jewish and Christian, deserved detailed record. The Christian claims must have seemed more extravagant to the pagans than even the Jewish ones. Thus Eusebius stated that the Christians were, in fact, born with the world, anticipating St. Augustine's vision of the city of God existing since the beginning of time.

In defending their religion against hostile critics, the early Christians were forced to fit some pagan history into their universal scheme. This was achieved by means

of universal chronologies from the creation of the world to each writer's own time. The events of Jewish and Christian history were thus synchronized with the main dates of the pagan myth and history. Sextus Julius Africanus, who wrote in the early 3rd century, is the first Christian writer known to have attempted this feat. He allotted 6,000 years to the whole span of human history and placed the birth of Christ in the year 5500 from the creation of the world. This work provided the model for the more elaborate Chronographia (Chronicle) of Eusebius. It became the foundation for a long succession of Greek chronographies produced by Byzantine writers. A Latin adaptation by St. Jerome (died 419/420) was immensely influential in western Europe for more than 1,000 years. A modern scholar is filled with mingled admiration and despair at the ingenuity of Eusebius and of his more eminent successors and at the absurdity of many of their conclusions. But they did originate and impose on the world a unified scheme of universal chronology. The dating from the birth of Christ was introduced by Dionysius Exiguus, who wrote at Rome in the early 6th century, and it was successfully popularized in the 8th century by the English historian Bede.

The writing of history of their own time was not an essential task for the Christians of the 4th and 5th centuries. When they did so, they wrote primarily in defense of their religion against the pagan world or against rival Christian groups branded as heretical. All these histories belong to religious apologetics. They suffer from inevitable distortions in the choice of what should be mentioned and what must be suppressed, and they are often excessively unfair to outsiders and opponents. These faults were not uncommon among the classical historians, though the Christians were somewhat unusual in their extreme conviction that they alone must be right. A comparison between the Christian historians and an outstanding pagan writer, such as Ammianus Marcellinus (second half of the 4th century), who was very ready to admire those Christians who merited it, brings out the intolerance and narrowness of outlook of his Christian contemporaries.

Eusebius was the earliest and the most important of the Christian historians of the 4th century. He is quite frank about the practical and apologetic aims of his

Historia ecclesiastica (written 312-324; Ecclesiastical History) designed to show how, through a long series of acts of Divine Providence, a Christian empire was finally brought into existence by Constantine. He admits that "we shall introduce into this history in general only those events which may be useful first to ourselves and afterward, to posterity." This work, like his other historical writings, is a mixture of devout fiction and invaluable detail. But there is plenty of the latter in Ecclesiastical History. Contrary to the usual practice of the ancient historians, Eusebius tries to specify his sources, and he quotes from them extensively in order to document as fully as possible the developments that resulted in the triumph of Christianity. He provided in this respect a valuable model for his medieval successors. The most astonishing thing about Eusebius was his capacity to handle his sources critically, in matters where it seemed permissible to do so. In one passage of his Chronicle he sets aside the authority of St. Paul in favour of a piece of evidence contained in the Book of Judges. In later patristic literature nothing similar is found.

Biography, as it was habitually written in antiquity, could be readily adapted to Christian purposes. St. Jerome modelled himself on Suetonius in compiling the lives of 135 Christian writers (written in 392) as a way of demonstrating the high level of culture attained by his coreligionists. The ancient biographers had freely mingled fact with fiction for the edification of their readers and could be readily imitated by the writers of the lives of Christian saints. The life of St. Anthony of Egypt by St. Athanasius (mid-4th century) set the pattern for this most popular type of medieval literature.

St. Augustine, the greatest of the Latin Church Fathers of the 4th and 5th centuries, was certainly not concerned with writing of history in any ordinary sense of the term. In his De civitate Dei (City of God) he might invoke historical evidence to demonstrate the utter degradation of all the non-Christian societies, and he encouraged his pupil Orosius to develop this theme more fully in the latter's Historiarum libri VII adversus paganos (Seven Books of History Against the Pagans, to 417). Nearly 200 manuscripts of Orosius have survived, testifying to the

immense popularity of his work in the Middle Ages. Augustine's greatest influence on historiography lay in his main message. His vision of the divine and the earthly cities confronting each other dominated the outlook of all the medieval Christian thinkers and profoundly affected their treatment of history. Within that divine plan for the world, purely secular history seemed an insignificant thing.

Early China.

The preservation of some records of historical events can be traced in China to at least the early part of the 1st millennium BC. Confucius (551-479 BC) was credited, rightly or wrongly, in the later Chinese tradition with editing the annals of his native state of Lu. But the appearance of the first works fully deserving the name of histories resulted from the unification of China under a single ruler in 221 BC. The first such work to survive, the Shih chi ("Historical Records"), dates from c. 85 BC. Its author, Ssu-ma Ch'ien, is quite justifiably called the father of Chinese historiography. His history exhibits many of the main features of the later Chinese official histories as they continued to be written down to the deposition of the last Chinese imperial dynasty in 1911. Within this fairly unified tradition, China produced a mass of historical writings unequalled by any other country before modern times. Until the late 19th century, Japanese historiography formed an offshoot of this tradition.

Chinese scholars showed an interest in the history of China from the earliest times. According to the Chinese conception, history makes sense only if it can furnish practical directives for action or supply correct information upon which action can wisely be based. All the schools of Chinese thought quoted the lessons of history. Confucius, with his stress on the moral content of these lessons, formed part of this universal belief in the value of history.

One of the duties inculcated by him was the scrupulous transmission of authentic records. When, some centuries after his death, the unified Imperial state began to recruit its bureaucracy among the Confucian scholars, the recording of all the necessary information and the careful preservation of records became one of the main functions of the Chinese government, both centrally and locally. A long

series of official histories and of records connected with them has survived from the time of the T'ang dynasty (618-907) onward. From then on, the great bulk of Chinese history was written by bureaucrats for bureaucrats. From a practical point of view this immense body of historical writings fulfilled a very useful purpose. Such histories were bound to be highly stereotyped and restricted in content to what interested the higher officialdom. It is easy to condemn it by modern Western standards for its excessive preoccupation with concrete details and inability to produce works of wider synthesis. But this Chinese tradition did gradually evolve in the direction of greater rationality and subtlety. Its scope widened as the sphere of government expanded. Furthermore, within this tradition there appeared from time to time writers of genius, men of bold critical spirit, genuine historical insight, and overriding integrity. One of the greatest was Liu Chih-chi (661-721), the writer of the Shih t'ung, the first thorough treatise in Chinese, or any other language, on historical method, which also constituted in effect a history of Chinese historiography. He had a successor in Ssu-ma Kuang (1019-86), the author of the first fairly comprehensive general history of China (covering the years 403 BC-AD 959). In the 17th century a remarkable group of historical scholars virtually founded a school of critical Chinese philology. None of these writers succeeded in radically transforming Chinese historiography, but they created an increasingly sophisticated and critical tradition. Their successors in the 20th century assimilated some valuable features of modern Western historiography.

MEDIEVAL HISTORIOGRAPHY

Europe from the 5th to the 11th century.

The period stretching from the 5th to the 11th century was a time of very profound cultural decline in regions that had once constituted the western half of the Roman Empire. Almost all the inhabitants of these provinces again became illiterate. There are long periods for which there are virtually no narrative sources, and the bulk of surviving historical writings consists merely of meagre factual annals. Virtually all the writers were ecclesiastics, in marked contrast to the Byzantine lands, where a

strong tradition of lay historiography persisted throughout the Middle Ages. The annalists and chroniclers of the West were predominantly monks, and their lack of experience of the secular world outside their cloisters made them into blinkered and unpractical historians. This was true even of Bede, an Anglo-Saxon monk, who was by far the greatest historian of the early
Middle Ages.

All the historians of this period were seriously affected by the cultural decline around them. They were having to write in part for a more uncultured audience. Sulpicius Severus, probably the best Western historian of the early 5th century, still intended his Chronica (to 403) for educated Roman Christians, but his life of St. Martin of Tours is a piece of medieval hagiography. This model could inspire lives full of folklore and miracle, from which the real human personalities of the saints were almost wholly absent. The same duality of purpose is a notable feature of Bede's voluminous writings. He explicitly recognized that he must adapt himself to his audience when he explained that he was writing in a simple Latin style so that he might be more easily understood by his Anglo-Saxonreaders. There is a marked contrast of tone between his theological and his historical writings. As a theologian, Bede follows Eusebius and the earlier Church Fathers in not exaggerating the frequency of miracles and in believing that they were most common in the earliest days of Christianity. But Bede's lives of the English saints and his Historia ecclesiastica gentis Anglorum (Ecclesiastical History of the English People), covering chiefly the years 597-731, are full of miracles and visions. There is one or other on almost every page. It is possible that some of these incidents were included by Bede because he thought that his readers expected mentions of these familiar, traditional stories.

In preparing his historical works, Bede not only took great care to assemble the widest possible collection of sources but also tells the reader what he is using. In dedicating his Ecclesiastical History to King Ceolwulf of Northumbria, he requests that in order to remove all occasions of doubt about those things I have written, either in your mind or in the minds of any others who listen to or read this history,

I will make it my business to state briefly from what sources I have gained my information.

An impressive list follows, including mentions of documents copied for him by friends at Rome, Canterbury, and other places. Like Eusebius, on whom Bede modelled himself, he quotes some of the documents integrally. Bede's methods of securing and recording information are so similar to the practices of modern historians and the judicious tone of his writing is so impressive that the reader is almost taken in into treating him as if he were a modern scholar. But Bede's Ecclesiastical History was written as a work of edification in order to strengthen the faith of his readers in Divine Providence, through which, as he saw it, his Anglo-Saxon countrymen had been converted to Christianity. All matters not connected with his main theme are ignored. Bede's handling of evidence on subjects that he regarded as embarrassing inspires mistrust. But these are small matters in comparison with the enormous mass of information that he alone has preserved and the encouragement that Bede continued to give for many centuries to the writing of history.

The influence of Bede and other Anglo-Saxon scholars was greatly felt during the later 8th and the 9th centuries in the Frankish kingdom, where under Charlemagne and his successor, Louis the Pious, there was a modest revival of historical writing. Besides the annals kept at various monasteries, which tended to convey information in a manner that suited the Frankish rulers, there were a few more ambitious ventures. The important Historia Langobardorum (History of the Lombards), written c. 774-785 by Paulus Diaconus, or Paul the Deacon, was the work of one of the best educated men of the time. Nithard, a grandson of Charlemagne, left an invaluable narrative of the disintegration of the Carolingian state during his lifetime. The work that exerted the greatest influence on the medieval writers of biographies was Einhard's Vita Karoli Magni (written c. 830-833; Life of Charlemagne). The author was a leading official and a close companion of Charles, and his work was naturally intended as a eulogy of the great king. Einhard says that Charlemagne retreated safely from Spain, returning with

his army safe and sound, except that on a ridge of the Pyrenees, on the way home, he happened to experience some small effects of Gascon perfidy. Nobody would gather from this that the Franks had narrowly escaped a major disaster. Einhard was merely echoing the story told in the semiofficial contemporary annals. Another source of distortion was Einhard's use of a classical model, the Lives of the Caesars by Suetonius. The subject headings under which he described Charles and even the very words used were partly borrowed from the lives of Roman emperors, but his Charlemagne is probably in essentials an authentic and credible portrait.

If bulk alone is to be taken as a criterion, annals were the main product of medieval historiography. The annalist merely sets down the most important events of the current year. In the case of the earliest medieval annals, the events were often noted down in Easter tables, in the blank spaces between the dates calculated for the forthcoming Easters. Such paschal annals would be extremely brief. When, as often happened, annals came to be written down in separate manuscripts, distinct from the Easter tables, there was room for the expansion of individual entries. In either case, the resultant annals cannot be regarded as history since the events are necessarily recorded in isolation. But they preserve in a right order the essential facts, which could be rearranged into a continuous narrative. Such a narrative, if it still followed the chronological arrangement of its various annalistic sources, should properly be termed a chronicle.

Medieval historians show little awareness of the process of historical change. They were unable to imagine that any earlier age was substantially different from their own. The unawareness of the meaning of anachronism helps to explain the strange wanderings of medieval annals and chronicles. If a religious community wanted to acquire a historical narrative, it copied some work that happened to be most readily accessible. A continuation might then be added at the manuscript's new abode, and, later on, this composite version might be copied and further altered by a succession of other writers. Hence there are at least six main versions of the annals known as the Anglo-Saxon Chronicle. They all derive from the annals kept down to 892 at

Winchester, the West Saxon capital. Thereafter, copies were acquired by religious centres in the most diverse parts of England, and one manuscript was being kept up to date at the abbey of Peterborough as late as 1154. An extreme case of wanderings is represented by the annals of the cathedral church of Cracow, the medieval Polish capital. The first section is based on Orosius, the next comprises annals beginning with the death of Bede and containing notices of Frankish and German events, while the Polish section starts with the conversion of Poland to Christianity (965-966) and ends in the 13th century.

Europe from the 12th to the 14th century.

Historians are accustomed to regarding the late 11th and 12th centuries as an age of intensified progress in culture and learning; this development, however, did not greatly affect historiography. There was a modest revival of interest in some of the ancient Latin writers, but would-be historians were unsure which ancient models they ought to imitate. A whole series of attempts was made to apply to other races the theme in Virgil's Aeneid of a noble group of people guided by the gods toward a splendid destiny. The first essential step was to establish the descent of one's nation from the ancient Trojans and then to trace subsequent history through a series of heroic conquests. The most ambitious of these writings was the Historia regum Britanniae (History of the Kings of Britain), by Geoffrey of Monmouth (died 1155), which attempted to establish for the Celts a historical destiny greater than any other. Although some, even contemporary, readers were not deceived by the work, and William of Newburgh, one of the best English historians of the 12th century, denounced it as a tissue of absurdities, many seriously accepted it as history.

With a few exceptions, the ablest minds of the 12th century were attracted into enterprises that ignored history; they were more concerned with systematization of thought and with philosophical speculations. One of the exceptions was Otto, bishop of Freising, in Bavaria. He was a grandson of the Holy Roman emperor Henry IV. He received the best education that his age could give, but he was also

briefly a Cistercian monk during the most austere period of that order's history. Otto was torn between conflicting impulses to seek the city of God as the only reality and yet to hope for the progress of the German empire. Out of this conflict came his first work, Chronica (The Two Cities), a chronicle of world history to 1146, perhaps the most profound medieval attempt at a Christian philosophy of history. As Otto himself confessed, it was composed "in bitterness of spirit . . . in the manner of tragedy." The election in 1152 of his nephew and friend Frederick Barbarossa, as emperor, filled Otto with a new elation. The excellence of his second work, Gesta Friderici I imperatoris (The Deeds of Frederick Barbarossa), derives in a considerable measure from a quality rare in medieval historians, a sense of optimistic belief in the value of writing history because it might become a record of human progress. The Deeds of Frederick Barbarossa contains a penetrating analysis of the problems encountered by the German rulers in trying to rule the precociously urbanized Italian society.

As in antiquity, the best medieval works were accounts of contemporary history by men who had participated in the events that they were describing. It is, however, very significant that some of the writers that are prized most highly today survive in only very few manuscripts and were presumably not appreciated by most of their contemporaries. One such work was the Historia pontificalis ("Pontifical History") covering the period 1148-52, of John of Salisbury, one of the most accomplished scholars of his age, who was writing about the period when he was in the papal service. Another instance of undeserved neglect is furnished by the Liber de regno Siciliae ("Book of the Kingdom of Sicily") covering the period 1154-69, written by an anonymous member of the Sicilian court.

Unlike the ancient historians, the medieval writers of contemporary history had no inhibitions about extensively quoting official documents. In England, a succession of writers preserved a large quantity of such texts. Roger of Hoveden was, in the last quarter of the 12th century, treated by the English kings as a kind of court historian. He preserved valuable legal and administrative records with which he was familiar through his activities as a royal official and justice. Matthew Paris, the

most important English monastic historian of the 13th century, was highly regarded by King Henry III and had excellent sources of information. He left behind a collection of transcripts of royal and ecclesiastical documents that today fills a large printed volume. Some writers made their chronicles into an anthology of official records, thinly connected by the author's brief comments. Such is the chronicle of Robert of Avesbury, consisting mainly of the military dispatches of King Edward III and other interesting documents to 1356. Another variant of the same method was for a wholly mediocre chronicle to incorporate exciting pieces of eyewitness narratives by other writers. A dull English monastic product of the late 14th century, the Anonimalle Chronicle, includes a narrative of the Peasants' Revolt of 1381, which is one of the most dramatic and interesting eyewitness accounts to be found in medieval historiography.

The most popular histories of the 13th and 14th centuries were encyclopaedic compilations giving all the important facts neatly arranged under the dates of popes, emperors, and other rulers. There were even more ambitious ventures aiming at summarizing all the important facts from all the different branches of human activity. The Dominican Order, created at the beginning of the 13th century, was especially concerned with producing such aids for the dissemination of useful knowledge. The best known of these Dominican works is the immense Speculum historiale ("Mirror of History"), by Vincent of Beauvais, written under the patronage of King Louis IX of France. It is a compilation made up of excerpts from many authors.

The 13th and 14th centuries were not a period of any fundamental innovations in the techniques and nature of historiography, but there was a growing diversity of types of historical writing. Very detailed, chatty narratives multiplied, often badly organized and inaccurate, but conveying the authentic atmosphere of the times and vividly portraying leading personalities. Such were the St. Albans chronicles of Matthew Paris (to 1259), the reminiscences of Joinville about St. Louis during the Seventh Crusade (1248-54), the Lombard chronicle of Fra Salimbene (to 1287), or the vast history of the first part of the Hundred Years' War written in the second

half of the 14th century by Froissart. Memoirs and histories written in vernacular languages, such as those of Joinville and Froissart, came to be quite common. Laymen began to write histories. Some were great men, like Geoffroi de Villehardouin, one of the leaders of the Fourth Crusade (which captured Constantinople 1202-04), of which he wrote an account. Important urban chronicles began to appear, such as the Florentine chronicle of Giovanni Villani, with its invaluable statistics of Florentine population and activities around 1338. The extraordinary personality of St. Francis, who died in 1226, inspired lives of him more convincingly human than any previous medieval biographies of saints.

The Humanist historians of the 15th century tried to make a deliberate break with the tradition of medieval historiography. By their insistence on a more coherent arrangement of subject matter, by their superior critical outlook, and, above all, by their much more accurate awareness of the process of historical change, they had introduced innovations of fundamental importance. In part they owed their grasp of these new possibilities to the influence of Byzantine scholars. In historiography, as in other matters, the new humanistic scholarship was a joint product of Western and Byzantine traditions.

BYZANTINE HISTORIOGRAPHY

During the millennium that elapsed between the collapse of the Roman Empire in the West in the 5th century AD and the Italian Renaissance of the 15th century, in no part of Europe did the writers of history consistently maintain as high a standard of achievement as in the Byzantine Empire.

Parts of the 7th and the 8th centuries form lengthy gaps in the record of Byzantine historiography, but this seems mainly to be the result of subsequent losses of manuscripts. When, in the middle of the 9th century, Photius, future patriarch of Constantinople, compiled a record of some 280 books that he had read, he mentioned works of 33 Greek historians, dating mostly from the late Roman Empire and the Byzantine period, 20 of which are now lost. But, among the

Byzantines of the 7th and 8th centuries, there was certainly no parallel to the Dark Ages in western Europe.

The Byzantine historians were heirs to the combined traditions of classical Greek writing, of the subsequent Hellenistic historiography, and of the Christian historical writing of the 4th century. Few ancient Latin historians were ever translated into Greek, and their influence on the Byzantines was, therefore, very slight. The older classical Greek historians provided the Byzantines with their cherished models of language and style. Like all educated Byzantines, the historians continued for a millennium to write in a literary language that soon became unintelligible to the vast majority of their compatriots. Hence, from the 6th century onward, there appeared, side by side with the learned historiography, a succession of popular chronicles written in the ordinary language. Most of these popular writings form - in their prejudice, ignorance, and crudity - a startling contrast to the works of the more eminent classicizing historians, but they do provide valuable glimpses of the sort of hagiographical history, more religious myth than sober fact, that ordinary Byzantines apparently wanted to read.

Herodotus and Thucydides were frequently invoked by Byzantine historians as models of fine prose. The influence of these two writers on the substance of what was written usually remained slight and superficial, however. The only Byzantine writers who seriously modelled themselves on these two oldest Greek historians wrote during the 15th century. The earlier Byzantine historians owed most to Polybius and to the Greek biographer Plutarch (died c. AD 119), the two Hellenistic writers who had the greatest influence on Byzantine notions of how history and historical biography should be written.

Like Polybius, the majority of Byzantine historians, including most of the best ones, perferred to write about their own times; and within these limits they produced some real masterpieces. Unlike the majority of the ancient historians, Polybius had included much autobiographical detail, and his influence reinforced the readiness of the Byzantine historians to talk about themselves, thus providing abundant information about several of these authors. Their histories are likely to be

one-sided and full of details about what interested them, while remaining silent about a great mass of other contemporary happenings. They are frequently gossipy and patently prejudiced, inspiring much less confidence than the austere, impartial writings of authors such as Thucydides. This is one of the main reasons why the Byzantine historians have often been excessively underestimated by modern readers. The bulk of the Byzantine contemporary histories were written by statesmen, high officials, and prelates - men with access to important information. They have to be used critically and cautiously but can be immensely valuable.

Priscus of Panium (c. 450), a member of a Byzantine embassy to Attila's camp, is the best source of information about that terrible king of the Huns and his followers. A century later, the reconquest of Vandal Africa and of Ostrogothic Italy by the emperor Justinian was the main theme of the History of the Wars of Procopius, a leading civilian adviser of Belisarius, the Byzantine commander. Subsequently, Procopius also wrote a Historia arcana (Secret History), containing a horrible indictment of the activities of Justinian and Belisarius. Many of his details about the corruption at court and the oppressive nature of the government may be substantially correct. In the 11th century Michael Psellus, who wrote a history of his own times, was a leading Byzantine scholar and official, for a time even the chief adviser of emperors. His Chronographia is concerned almost entirely with the happenings at the Byzantine court and is one of the most gossipy and amusing narratives ever written on such a subject. His psychological insight and his lively and subtle style delighted the educated Byzantines. Anna Comnena, the daughter and biographer of the emperor Alexius I, greatly admired Psellus. Her own Alexiad is a much less fascinating work, but the recovery of the Byzantine power under her father provided her with an important theme.

The last, increasingly disastrous, centuries of Byzantine history are recorded by a series of scholarly and interesting historians. Nicetas Choniates, a high imperial official, provides a surprisingly balanced eyewitness account of the siege and capture of Constantinople by the forces of the Fourth Crusade (1202-04). George Acropolites, a leading adviser of the Greek emperors of Nicaea, carries the story

from 1203 to the recapture of Constantinople by the Byzantines in 1261. The later 13th and 14th centuries are covered by a succession of writers deeply immersed in contemporary theological disputations. Perhaps the most readable of all Byzantine histories is the largely autobiographical work of the leading politician and emperor John VI (reigned 1347 to 1354), written after his deposition during his years of enforced retirement in a monastery. George Sphrantzes, a close friend of the last emperor, Constantine XI, included in his history an eyewitness account of the siege and capture of Constantinople by Mehmed II in 1453. Two of Sphrantzes' contemporaries chose to write primarily about the Turks. Their methods place them among Renaissance historians. Laonicos Chalcocondyles wrote (in about 1464) an account of the rise of the Turkish state. He did so in the manner of Herodotus, with long digressions on various neighbouring nations. A little later, Critobulos of Imbros, in his account of the Turkish conquest of Constantinople, made Mehmed II his chief hero and modelled his history on Thucydides.

The study of what might be called "historical antiquities" was not much cultivated by Byzantine scholars. The most notable exception was the emperor Constantine VII, but only some fragments of his voluminous collections have survived (dating from about 940 to 959). They include a very interesting account of the various peoples with whom the Byzantines had to deal. Such ancient Greek literature as still survives, including that of all the historians, was preserved by the Byzantine scholars. When, around the year 1400, the teaching of Greek was introduced into Italian universities by Byzantine scholars, they brought also their superior techniques of literary scholarship, transforming thereby the study of Latin authors as well as introducing into western Europe the treasures of Greek literature. One result was the emergence of the new Renaissance historiography.

MUSLIM HISTORIOGRAPHY

Muslim historiography appears to have originally developed independently of European influences. Until the 19th century Muslim writers only very seldom consulted Christian sources and almost never noted events in Christian countries.

Fortunately, they displayed at times more curiosity about the non-Muslim peoples of Asia. The first and best history of the Mongol conquests in the first half of the 13th century was the work of a Persian, Joveyni. On a visit to Mongolia in 1252-53, he was able to consult the recently compiled, earliest Mongol narrative (Secret History of the Mongols).

The origins of Arabic historiography still remain obscure because of the gap between the legendary traditions of pre-Islamic Arabia before the start of the Muslim era (AD 622) and the sophisticated and fairly exact chronicles that began to appear in the later 8th and 9th centuries. But while the detailed stages of this development still await reconstruction, the main influences shaping the early Muslim historiography are clear enough. As in the case of the ancient Jews, it was created and perpetuated by religion. Muhammad (died 632) regarded himself as a successor to a long series of Jewish and Christian prophets, and he made Islam a religion with a strong sense of history. The Qur`an, Islam's holy book, is full of warnings derived from the lessons of history.

Teachings of Muhammad not included in the Qur`an came to be regarded after his death as authoritative tradition left behind by him. All his sayings and actions were therefore carefully treasured and ultimately came to form, in combination with the Qur`an, the foundation for the body of Muslim law (Shari'ah), common to all Islamic communities. These traditions (Hadith) were transmitted orally for several generations, until they were written down in the 8th and 9th centuries. The resultant collections were only partly historical, as myths and inventions crept into them. The scholars who were engaged in preserving and verifying these traditions were chiefly preoccupied with organizing them into legal and theological systems, and they were frequently hostile to the historians. The earliest authoritative life of Muhammad, written by Ibn Ishaq (died 768), was attacked by a leading exponent of the legal "traditionist" learning. This confirms the independence of the historical scholars from the theological and legal interests. But both groups shared some common materials, and the strict rules evolved by the legal "traditionists" for recording their sources and tracing a continuous chain of authoritative transmitters

of the traditions encouraged similar exact habits in the Muslim historians. The resultant histories were often pedantic, full of unrelated facts, and deficient in reflective comment, though there are some astonishing exceptions, such as the writings of Ibn Khaldun (1332-1406). But the better Muslim historians scrupulously quoted their authorities and tried to be truthful. This was particularly true of the "classical" school of historians, who were writing at the centre of the 'Abbasid caliphate in Iraq in the 9th and 10th centuries. At-Tabari (died 923), the most authoritative of them all, wrote his "History of Prophets and Kings" as a supplement to his earlier commentary on the Qur`an, and subsequent Muslim historians were content to follow his reconstruction of the early Islamic history. The Syrian and Iraqi historiography of the 12th and early 13th centuries is at least as valuable as the Western historical writing of this period, and sometimes it is clearly better.

To orthodox Muslims, the development of the Islamic community represented a continuous manifestation of God's purpose. Consequently, the recording of the religious progress of the Islamic society continued to be sacred duty. One of the original features of Muslim historiography is the large amount of attention devoted to the lives of devout men and of scholars. To many Muslim historians, these spiritual and intellectual activities were of much greater importance than the doings of princes and warriors. One of the peculiarities of Muslim historiography was the liking for encyclopaedic dictionaries of famous men. The earliest of these were devoted to the Companions of Muhammad and to the early transmitters of the Muslim traditions. For a thousand years extremely diverse types of biographical collections have continued to appear in the Muslim world. Those devoted to religious scholars attained a particularly wide diffusion. Saladin (Salah ad-Din), who took Jerusalem from the crusaders in 1187 and later opposed the Third Crusade, offered to the Muslim writers the particularly congenial subject of a ruler dominated by a sense of religious duty. A particularly fine example of medieval Muslim historiography is the biography of Saladin by Baha` ad-Din (died 1234),

which gives an exceptional insight into Saladin's motives for many of his critical decisions.

But the greatest Arab historian and one of the most penetrating thinkers about historiography in any time or place was undoubtedly Ibn Khaldun. The introduction (al- Muqaddimah) to his Kitab al-'ibar, a universal history (begun in 1375), is, in A.J. Toynbee's judgment (1934), "the greatest work of its kind that has ever yet been created by any mind." Ibn Khaldun had absorbed all the learning accessible to a Muslim of his time. He was a master of religious learning, an outstanding judge, a writer on logic. He turned a subtle and most disciplined mind to historiography in order to explain his personal tragedy. He had served a succession of rulers in Islamic Spain and the Maghrib (Northwest Africa) as a general, a politician, and even once as a chief minister, and his activities had always ended in disaster. In order to explain what had gone wrong, he sought to achieve a correct understanding of the forces that governed the societies known to him. He concluded that political stability had become impossible in his native Maghrib, because over centuries economic prosperity had declined excessively and the forces of lawlessness had become too strong.

As a detailed chronicler of events Ibn Khaldun is not always exact, but, like contemporary historians, he knew how to reconstruct correctly the main trends over several centuries. His ability to formulate general laws that govern the fate of societies and to establish rules for the criticism of sources provided him with an intelligent framework for the correct reconstruction of past history.

Ibn Khaldun's Muqaddimah has survived in at least a score of manuscripts, but he has had no effective influence on Muslim historiography until recently; after his time, as before, the writing of history continued to be a normal feature of Muslim civilization in the more advanced Islamic societies.

In several countries, notably in parts of India, the first works that deserve the name of history appeared only after the Muslim conquest or the conversion to Islam. After the 12th century Arabic ceased to be the main language of Muslim historiography. Distinguished histories were written in Persian in the 13th century,

and subsequently Turkish and other vernaculars came to be used by historians in different parts of the Islamic world. But, in its isolation from non-Muslim influences and its traditional interests, Islamic historiography underwent no intrinsic change until the 19th century, when it began to be affected by the impact of modern Western civilization.

The early Humanists.

If there is one thing that united the men of the Renaissance, it was the notion of belonging to a new time. Lorenzo Valla, one of the ablest of the early Humanists, in a preliminary draft of his history of King Ferdinand I of Aragon (written in 1445-46), proudly enumerates the modern technical inventions made in recent centuries, and especially near his own day. The sense of the novelty and excellence of their achievements was particularly felt by the men of the Renaissance in connection with their attempts to imitate the works of the ancient Greek and Roman writers and artists. They were not yet claiming that an era of unlimited progress was dawning for mankind - such concepts belong to the 18th century - but the belief in the progressiveness of their own age soon spurred the best Renaissance scholars and artists into achievements that, in some important respects, surpassed their ancient models. This happened in historiography, and especially in the sciences connected with it. The pace of change must not be exaggerated, however. Despite promising beginnings, historiography as a systematic discipline did not emerge during the Renaissance and, in fact, this development did not occur until the 19th century. The reasons for this delay form one of the main problems in any study of historiography between the years 1400 and 1800.

In the early Renaissance one by-product of the newly won sense of modernity was the tendency to regard the millennium between the collapse of the Roman Empire in the West and the 15th century as an era of prolonged decline. The concept of the Middle Ages was thus introduced for this intervening period. Two very important histories written in the first half of the 15th century deliberately concentrate on the

medieval centuries. Their authors were leading Italian Humanists. The first to appear was the Historiae Florentini populi ("History of Florence") of Leonardo Bruni, the city's chancellor from 1427 to 1444. The second, the Historiarum ab inclinatione Romanorum imperii decades ("Decades"; mainly devoted to Italy), was written by Flavio Biondo, an important papal official. It covered the period from the sack of Rome by Alaric in AD 410 to the writer's own time. The "invention" of the Middle Ages as a separate historical period remains one of the most enduring legacies of Renaissance historiography.

Unlike the medieval historians, the Renaissance Humanists became much more acutely aware of the process of historical change. This was a gradual development. They were trying to understand the ancient writers, whom they were seeking to emulate, and they became increasingly aware of the need to replace these writers in their correct historical setting. When Petrarch (1304-74), the pioneer Italian Humanist, unearthed in 1345 a collection of Cicero's letters, he was shocked to discover that Cicero was not a cloistered scholar of the medieval tradition but a busy politician who wrote his dialogues in moments of banishment from active life. In 1361, in a letter to the Holy Roman emperor Charles IV, Petrarch was able to use his increased familiarity with classical documents to expose a medieval forgery of the Austrian archduke masquerading as a charter of Julius Caesar.

Between about 1440 and his death in 1457, Valla was one of the most influential Humanists. His Elegantiae linguae latinae (1444; "Elegancies of the Latin Language") was a treasury of information about correct Latin usages. For Valla the meaning of words was not natural but conventional and historical, because it was derived from changing custom. Thus a sense of ceaseless historical evolution was planted at the very centre of Humanist preoccupations with the recovery, the correction, and the interpretation of ancient texts.

In 1440 Valla's patron, King Alfonso of Naples, at war with the papacy, asked Valla to write a treatise against Pope Eugenius IV. Valla obliged by decisively disproving, on both linguistic and historical grounds, the genuineness of the "Donation of Constantine." From the middle of the 8th century, when this

document was probably concocted, it had been used by the popes as one of the weightiest justifications for their claims to secular authority in Italy. Its authenticity had been sometimes questioned in the past by some of the acutest minds, such as Bishop Otto of Freising in the 12th century and Marsilius of Padua in the first half of the 14th century, but it required Valla's expert techniques to dispose of the "Donation" forever. The validity of Valla's methods of historical criticism was at once recognized by at least one other leading Humanist. Biondo wrote the relevant portions of his "Decades" of papal and Italian history between 1440 and 1443, while remaining in the service of the very same Eugenius IV who had been the chief object of Valla's attack. Yet Biondo tacitly accepted Valla's conclusions, and he never mentions the "Donation of Constantine." Biondo's critical outlook found still another expression in his summary dismissal of the fabulous history of Geoffrey of Monmouth. In his copy of Geoffrey he entered only a single note: "I have never come across anything so stuffed with lies and frivolities."

Historical philology.

Valla's work on the texts of the New Testament proved in the long run to be one of the most influential applications of the new science of historical philology. His aim was to recover, so far as possible, the original Greek version through the use of the oldest extant manuscripts. He defended these researches by pointing out that he was not correcting the Holy Scriptures but merely the Latin Vulgate translation of St. Jerome that had been adopted by the Catholic Church. The revolutionary nature of Valla's historical approach comes out most strikingly in his comment that "none of the words of Christ have come to us, for Christ spoke in Hebrew and never wrote down anything." The corrections assembled by Valla became generally known when, in 1505, Erasmus published them as Annotationes on the New Testament. They provided a model for Erasmus' edition of a Greek New Testament in 1516, from which stem all the new Protestant versions of the 16th century.

The new historical philology was also soon applied to the study of philosophical and legal texts. In this, the most striking progress was made in the second half of the 15th century by Politian, who lectured at Florence, and by his friend Ermolao Barbaro, who taught at Padua. They were inaugurating the history of ideas and of intellectual movements. In his studies of Aristotelian texts, Barbaro insisted on using only the commentators of antiquity. In his lectures and writings (1489-94), Politian tried to reestablish from internal evidence the correct sequence of Aristotelian treatises, and he traced the gradual liberation of Aristotle's thought from the influence of Plato. The meaning of the terms used by Aristotle was rigorously investigated in the light of the linguistic usage of his Greek contemporaries. Politian's ventures into the field of legal texts proved particularly influential.

He had at his disposal a very good 6th-century version of the Digest - that is, the section of Justinian's Corpus Juris Civilis (Body of Civil Law) based on the rulings of the Roman jurists. Politian's collation of it with the first printed edition of the Digest (in 1490) formed part of an inquiry into the transmission of the texts of the Roman law during the Middle Ages. Politian's researches stimulated a remarkable school of Humanist jurists, mostly Frenchmen, headed by Guillaume Budé, who published the first historical commentary on the Digest in 1508. In the course of the 16th century, these scholars laid the foundations of a new branch of scholarship, the history of laws and institutions.

The methods of textual criticism used by Politian and his friends were designed to produce definitive editions of classical texts. Politian was aware of the need to establish the correct descent of manuscripts and to disentangle the best textual tradition. In all this he was far ahead of almost all his contemporaries, and he was anticipating the procedures that were systematically adopted for the first time by Karl Lachmann and other German scholars in the 19th century. The historical philology of Politian was a program for the future rather than a dawn of a new era in the editing of classical texts. In contrast to his methods, most of the other Humanist editions of the Latin and Greek classics are very unsatisfactory. This is

particularly true of the editions produced between about 1400 and 1550. The reckless emendations of Humanist editors, coupled with the subsequent disappearance of some of the manuscripts used by them, created grave problems for later scholars. Ever since the 17th century the task of the more modern editors has consisted largely in reconstructing, so far as possible, the manuscript versions available before 1400.

Notable works from the period.

Modern historiography was created in the 19th century through a successful combination of the use of narrative sources with every other type of evidence. Some 15th-century Italian Humanists were already aware of these possibilities. The idea of recovering an entire civilization through a systematic collection of all the relics of the past was not alien to them. Biondo used mainly conventional narrative sources for his "Decades" of Italian history, but his description of the city of Rome in antiquity (Roma instaurata, 1444-46) was based on a novel combination of the narratives of other historians with a wide range of miscellaneous sources. These included topographical guides, public and private documents, studies of surviving buildings, inscriptions, and coins. But in practice most histories and biographies continued to be written in a conventional way, while the revived study of "antiquities" was cultivated in separation from narrative historiography.

Imitation of ancient models is the feature most often stressed in the modern descriptions of Humanist histories. This meant that style mattered at least as much as content and that historical truth might be obscured by literary conventions. On the more positive side, there was the renewed insistence on the choice of definite, clearly delimited subjects and on a more coherent arrangement of material. The abler Humanist historians, however, were also making innovations that bring their practice a little nearer to present notions of writing history.

Several Humanist historians were particularly attracted to the study of the origins of the states about which they were writing. In the 15th century Bruni did this for Florence, and Biondo and Bernardo Giustiniani for Venice, to mention some

notable examples. In the 16th and early 17th centuries, French and English scholars inaugurated a critical study of the origins of their national institutions. Humanist historians prided themselves on their critical ability to overthrow the legends in which various countries had concealed their ignorance of their own origins. The incentives to revise the earliest history were often political. Bruni deemed it essential to prove that Florence had not been founded under the tyranny of the Roman emperors but in the time of the free republic. He happened to be right. The Humanist historians were more confident than their ancient predecessors that they could write competent histories of a remote past. In practice they were much less successful in this than they imagined. In dealing with periods before their own time, they usually followed only a restricted number of earlier narratives, though the best of them, such as Bruni and Biondo, displayed in their histories of medieval Italy a novel ingenuity in combining well-chosen sources. Biondo, for example, made effective use of Dante's correspondence.

There was also some modest progress through the better use of documentary sources. This is often far from obvious, because Humanist historians, like their ancient predecessors, do not usually refer to their sources, even when they quote texts verbatim. Hence came Leopold von Ranke's utter misjudgment of the historical value of the Storia d'Italia ("History of Italy") of Francesco Guicciardini. Before Ranke's time it was universally accepted as the most authoritative contemporary history of Italy in the years 1494 to 1534. Ranke, who became one of the pioneers of "scientific" history in Germany, first established his reputation in 1824 by his attack on the reliability of Guicciardini. Ranke argued that the statements of that great Florentine statesman were contradicted by documentary evidence and that his history must have been based on unreliable secondary authorities. The discovery in the 20th century of Guicciardini's private archive proved that his history was scrupulously based on original documents of the highest value.

Guicciardini, in a work that forms the nearest Renaissance parallel to the history of Thucydides, tries to comprehend the succession of tragedies that befell Italy from

the start of the French invasions in 1494. This desire to recapture the rational causes of events is one of the most mature features of the best Renaissance historiography.

EARLY MODERN HISTORIOGRAPHY

Spread of Humanism.

Italian Humanist historians provided models that could be imitated easily in other countries. Almost everywhere in western and central Europe, local writers were encouraged to produce descriptions and histories of their own lands, intent with patriotic pride. In such countries as Spain and Poland, which had only recently achieved their unity, this was a way of commemorating their newly won cohesion. In the 15th century it was the object of a pioneer work on the earliest antiquities of Spain, the Paralipomena Hispaniae, by the Catalan Humanist bishop Joan Margarit i Pau, and of the invaluable Annales seu cronicae incliti regni Poloniae ("History of Poland"), by Jan Dlugosz, which included an exceptionally precise geographic description of his country. In Germany a sense of national identity could be vindicated by Humanist historians striving to minimize the importance of the continued political division of their land. The Germania of Tacitus was printed in Germany as early as 1473 and started the fashion of using this collective name for that country. Tacitus called the Germans "the indigenous inhabitants." This was used by a leading patriotic Humanist, Conradus Celtis, as a proof that Germany should be free from all foreign domination. Celtis and his other Humanist contemporaries deliberately hunted for manuscripts of medieval German writers to prove that their country, despite its disunity, could have a national history. Some important masterpieces were recovered, including the histories of Otto of Freising. Celtis' pet project of a description of Germany modelled on Biondo's Italia illustrata was carried out in 1530 by Sebastian Münster, and Münster's fuller Cosmographia (1544; "Cosmography"), though purporting to describe the known world, devoted one-half of its 818 pages to "the German nation." There was also a spate of histories of Germany, mostly very laborious and unreflective but

incorporating the newly rediscovered medieval narratives and even some documentary sources. Greater originality came only in the wake of the Reformation. The same thing happened in France and in England. In both countries patriotic preoccupations were a leading feature of works written by Humanist historians, and the appearance of Protestantism reinforced in a peculiar way the existing nationalist tendencies.

The influence of the Reformation on historiography must first be discussed at a more universal level. As the philosopher Francis Bacon shrewdly observed, Martin Luther had been obliged "to awake all antiquity and to call former times to his succours so that the ancient authors which had a long time slept in libraries, began generally to be read." This was not because Luther would have regarded himself as a historian. But as early as 1519, in his disputation with Johann Eck, he encountered the assertion that the primacy of the pope was of divine origin. In order to disprove this and to demonstrate that they alone represented the true church, the Protestants had to retell in a new way the entire history of Christianity. In a preface to the Vitae Romanorum pontificum ("Lives of the Pontiffs"), published by Robert Barnes in 1535, Luther himself confessed that, although he himself had not originally attacked the papacy with historical arguments, now it is a wonderful delight to me to find that others are doing the same thing from history - and it gives me the greatest joy to see that history and Scripture entirely coincide in this respect.

Protestant history.

The starting point for the Protestant rewriting of Christian history could best be found in St. Augustine's teachings. The true church, the city of God, had always existed, even though at times it seemed to be overshadowed by the enemies of the divine order. Those enemies were not only the pagans and the heretics, as St. Augustine had believed. In more recent times they had included also the upholders of the papal authority and the persecutors of such medieval true Christians as John Wycliffe (died 1384) and John Hus (died 1415). The writings of Eusebius provided

the model for chronicling the sufferings of the faithful until the dawn of freedom for the true church in the 16th century. These views about the correct history of Christianity were presented with exceptional cogency in John Calvin's Christianiae religionis institutio (fullest edition 1559; Institutes of the Christian Religion) and were shared by most Protestant scholars. The only obvious disagreements arose when Protestants tried to pinpoint the moment at which the church took the fatal turn away from God's true purpose. While the radical sectarians considered that the papacy had always been corrupt, less extremist Protestants were prepared to accept the earlier popes and to argue that the rot set in at some date between the time of Eusebius (died c. 340) and the 7th century. The choice of precise date might depend on the national traditions of each country. Thus, Bishop Richard Davies, in his preface to the New Testament in Welsh (1567), treats Pope Gregory the Great (died 604) as a special enemy because Gregory's effort to convert the Anglo-Saxons led ultimately to the subjugation of the autonomous British church.

Historians writing in this spirit were incapable of impartiality. But the historical controversies between the Catholics and the Protestants produced from both sides huge compilations. Their authors were determined to prove their respective cases by a stupendous marshalling of authorities and documentary sources. The habit of giving copious references and long, exact quotations, missing from the Humanist historiography, was reintroduced by the religious controversialists. On the Protestant side, the largest work is the Ecclesiastica historia, or the so-called Centuriae Magdeburgenses (13 volumes, 1559-74; "Magdeburg Centuries"), retelling the history of the church down to 1200. The Catholic reply, equally huge and graceless, was produced in 12 volumes by Cardinal Baronius. The chief Protestant critic of this work, the great Greek scholar Isaac Casaubon, was astonished by the Cardinal's ignorance of Greek and Hebrew, his gross mistakes, and his boundless credulity.

The narratives of contemporary events written in the 16th and early 17th centuries by the participants in the religious struggles, though equally partisan, include some works of great historical value and high literary merit. The earliest and best

German Protestant narrative, that by Johannes Sleidanus, received a grudging tribute from his great opponent, the Holy Roman emperor Charles V, who remarked that "the rogue has certainly known much; he has either been in our privy council or our Councilors have been traitors." John Foxe's Book of Martyrs (1563) contains a great mass of exact information about the persecution of reformed religion in England and Wales during the reign of Mary Tudor, and it has influenced many generations of British Protestants. The achievements of Queen Elizabeth I and the Anglican Church's settlement of her reign found an outstanding defender in William Camden, who was encouraged to write by Elizabeth's leading ministers. In his Annales Rerum Anglicarum, et Hibernicarum Regnante Elizabetha ("Annals of Elizabeth's Reign") Camden made excellent use of a mass of official records at his disposal, though his treatment of confidential matters had to be discreet.

Out of a conflict between Venice and the papacy in the first years of the 17th century was born the Istoria del concilio tridentino (1619; History of the Council of Trent, 1676) of Fra Paolo Sarpi. A Catholic friar, but a passionate defender of Venetian autonomy, Sarpi drew a dark picture of worldly papal policies and the unscrupulous machinations of the Jesuits. It is a bitter, prejudiced, but splendidly written and well-informed work, which profoundly influenced the anticlerical historians of the 18th century. All these contemporary narratives, however, have one serious limitation. They deal almost exclusively with political events and with changes in ecclesiastical organization. The Protestant schism is treated as merely a revolt against the abuses of the old church, and the deeper reasons for the alienation of the Protestants from the Catholic faith are never explained. Furthermore, these historians, by attributing the origins of the schism almost exclusively to Luther's sudden conflict with the papacy, obscured the existence in the early 16th century of numerous Catholic reformers, whose sole aim was to transform the Catholic Church from within.

This one-sided approach to the history of the Reformation was destined to persist for a long time. Two influential histories published in the years 1683-88, one by a

great Catholic prelate, Bishop Jacques-Bénigne Bossuet, and the other by Pierre Jurieu, a leading Protestant, still agreed on the same superficial account of the causes of the Reformation.

The rewriting by the Protestants of universal church history naturally involved a drastic revision of the history of the national churches. In Germany, particularly, the history of the church had become inextricably intermixed with the destinies of the German empire. Their hatred of the papacy made the Lutherans visualize the course of German history with unusual clarity. Nobody before them had attempted to impose on that history a single intelligible pattern of any sort. Theirs was bound to be a prejudiced pattern, a story of gradual national disintegration as the result of the successive defeats of the German emperors by the papacy. Johannes Stumpf's tragic chronicle of the Holy Roman emperor Henry IV (published in 1556) treated his struggles with Pope Gregory VII as the beginning of the empire's tribulations. The whole course of German history was retraced in this fashion under the influence of Luther's chief Humanist collaborator, Philipp Melanchthon, in the so-called Chronicle of Carion, written in its final versions (1572-73) by Melanchthon's son-in-law, Caspar Peucer.

One of the most novel features of the English Protestant historiography was the reawakening of scholarly interest in the period before the Norman Conquest of England in the 11th century. Matthew Parker, Queen Elizabeth's first archbishop of Canterbury, thought he could discern in the pre-Conquest church elements of true Christianity that were destroyed thereafter and had only been reintroduced by the Protestants. The Anglican Church could be represented as a return to the traditional practices and beliefs of the early English Christians. Thus the replacement of Latin by English in the Protestant church services could be justified by citing the presence in Anglo-Saxon England of Bibles, liturgies, and devotional literature in the Old English language. Parker and his friend Lord Burghley, Elizabeth's most trusted minister, gathered around them a circle of enthusiastic scholars, whose work preserved most of the important Anglo-Saxon texts as well as of some leading post-Conquest chronicles. Parker's own method of editing texts horrifies

modern scholars, but some of the antiquarian works published by members of this group were of high quality.

Camden's Britannia (first edition 1586, later much enlarged) was a pioneer work on the topography of Roman and early medieval Britain. The edition by Sir Henry Spelman of the records of the pre-Conquest church councils was the first serious attempt to apply to an important type of early sources the best methods of continental scholarship.

Historical outlook and legal histories.

The growth of a historical outlook can be traced in the 16th century in many diverse fields of learning. For the first time men were realizing that there was a historical side to every branch of knowledge concerned with human affairs. "I have become aware that law books are the products of history," wrote the French legal historian François Baudouin in 1561. In each branch of study there developed a special historical technique particularly appropriate to it. The most sophisticated scholarship was to be found in the field of classical studies. A group of scholars active in the second half of the 16th century were achieving results much superior to the work of the earlier Renaissance classicists. They combined philological expertise with a determination to reach a really adequate understanding of the ancient Greek and Roman civilizations. A few were Italians, such as Carlo Sigonio, but most of the important works were written in France and in the Protestant centres of Switzerland and Holland. As textual critics these scholars were reacting sharply to the earlier, more haphazard, methods of emending and editing classical authors. They were trying to bring the text of one writer after another to a state of near perfection. Some leading ancient historians, such as Tacitus, benefitted greatly from this treatment (edition of Lipsius in 1575). Though their methods do not quite reach the standards of modern scholarship, they anticipate intelligently many of the procedures more systematically adopted in the 19th century. Isaac Casaubon was the first to point out in his edition of Suetonius (1595) that Einhard's 9th-century life of Charlemagne was modelled on the work of that Roman historian.

Casaubon's friend Joseph Scaliger renewed the science of classical chronology (1583) and was the first to reconstruct the original Greek Chronicle of Eusebius lying behind St. Jerome's Latin translation. Sigonio's pioneer work on the rights and duties of Roman citizens (1560) was later much used by Theodor Mommsen, one of the founders in the 19th century of the modern study of Roman history.

In the course of the 16th century, non-narrative historical work of the highest originality and complexity was being carried on in the legal faculties of French universities. One important stimulus was provided by the existence in France of different legal systems - the uncodified provincial customs in the north and the written law in the south. The latter ultimately derived from the Roman law, and, in the southern French universities, there arose an eager demand for the introduction of the new Italian methods of interpreting the Roman legal texts. Andrea Alciato, a pioneer in the historical treatment of the Roman law, taught at Bourges from 1529 to 1533, and his pupils founded the "Romanist" school of French legal historians.

Important advances were made in the study both of the Roman law and of the origins of the French legal customs, laying virtually the foundations of a new branch of scholarship, the history of law and institutions. François Baudouin published in 1545 the first historical survey of the development of the Roman legal science. The treatise on the custom of Paris by Charles Dumoulin (published 1539-58) resulted from his advocacy of the codification of the northern French legal customs. It was the first scholarly exposition of a body of customary French law derived from feudal practices, and it amounted to a first comprehensive history of European feudalism. It prompted a series of controversial works by a succession of scholars. The Roman, the Germanic, and the Celtic roots of feudalism all found advocates, and the respective claims of Lombard and Frankish texts to provide the best clues were vigorously canvassed. The complexity of the problems presented by the unravelling of the origins of feudalism dawned on scholars for the first time. The most valuable of these attempts to rediscover the "ancient French constitution" were the researches on "the antiquities of France" of Étienne Pasquier (published 1560-1607), which form a basis for all later study of medieval French institutions.

Secularization.

One of the novel features of European civilization in the later 16th and 17th centuries was a secularization of mental interests. Secular learning could now produce ideas more fascinating to intelligent men than theology. History was one of the most popular types of literature sought by a growing reading public. Several treatises on the proper way of writing history appeared in the third quarter of the 16th century. An anthology consisting of 12 such works, including the famous Methodus of the French political philosopher Jean Bodin, was published at Basel in 1576. Nearly 100 years later a "Catalogue of the Most Vendible Books in England" (1657) showed that history books constituted a large proportion of the total works published. It has been estimated that between 1460 and 1700 at least 2,500,000 copies of 17 leading ancient historians were published in Europe.

The late 16th century and the 17th witnessed the publication of several great collections of historical materials. The men who undertook these gigantic tasks often were antiquarians accumulating miscellaneous records rather than historians, but they were supplying materials for generations of future historians. Some of the most important publications of sources appeared in France and the Netherlands. Pierre Pithou was a pioneer in editing materials for the history of the Frankish period. The collections of André Duchesne are a vast storehouse of chronicles and other sources for the study of medieval French history. Le Nain de Tillemont edited 20 volumes of records devoted to Roman and church history during the first six centuries of the Christian Era, which a century later furnished one of the principal sources for Edward Gibbon's work The History of the Decline and Fall of the Roman Empire. In 1629 a Belgian Jesuit, Jean Bolland, embarked systematically on the editing of records connected with all the saints whose feasts had at any time been celebrated by the church, and this series of publications has been continued to the present day. In the second half of the 17th century, the French Benedictine congregation of Saint- Maur started an immense series of publications commemorating the history of the Benedictines and of other monastic

orders. The greatest Maurist scholar, Jean Mabillon, was accepted throughout Europe as the most erudite historian of his time.

In spite of its popularity among an expanding reading public and of the large number of learned editions of materials that it inspired, history was not, for most of the 17th century, one of the sciences that made men proud of living in a modern age. Immense progress was taking place in mathematics, astronomy, and physics. History not only did not seem capable of much further development, but scientifically minded men were beginning to dismiss it as a branch of knowledge that would never be worthy of serious respect. Mabillon's De Re Diplomatica (1681) helped to challenge this pessimistic view, but a further century elapsed before history began to be accepted as an authoritative discipline.

One major obstacle to the progress of historiography was the hostility of rulers to publications that did not favour their governments. The growth of an influential reading public made rulers increasingly suspicious of historical writings; for example, the censorship exercised by Cosimo I de' Medici, ruler of Florence from 1537 to 1574, precipitated the decline of Florentine historiography. Comparisons with the past also could be invidious.

In 1599 Elizabeth I of England censured an author for describing the deposition of one of her predecessors, Richard II, 200 years earlier. Fear of possible trouble made highly intelligent scholars into one-sided historians. The great jurist Hugo Grotius avoided in his history of the wars of the Dutch against Spain discussions of the religious aspects. Samuel Pufendorf, the historian of the Swedish conquests, carefully left out the internal developments in 17th-century Sweden.

Bacon, Descartes, and Mabillon.

The scholars who in that century were responsible for the great advances in the mathematical sciences were convinced that their achievements would ultimately give mankind a novel mastery over its natural environment. This is particularly true of Francis Bacon and of René Descartes.

Their optimism was laying the foundations for a belief in a possibility of continuous progress without which the purposeful and assured historiography of

the 19th century would be inconceivable. But the attitude toward history of most of the leading thinkers and scientists of the 17th century was not helpful to its immediate development. Bacon, who wrote a readable and rationally argued biography of King Henry VII of England, attached no importance to accuracy; for example, he antedated Henry's death by a whole year and could not be bothered to undertake any detailed research. Gottfried Wilhelm Leibniz was a great mathematician, but his attempts to apply science to historiography led to mechanistic constructions from which real human beings were largely missing. Numerous influential thinkers were decidedly hostile to history. Descartes, the most eminent of the anti-historical scientists, was not simply disgusted by the unsystematic and imprecise methods of the historians of his time but also doubted whether, strictly speaking, history could be regarded as a branch of knowledge at all. But it is important to remember that much of the 17th-century criticism of history was an attitude of men who simply had other priorities and were concerned to attack doctrines that, for one reason or another, historians seemed to support. In the late 17th century the most successful defenders of history were the members of certain particularly scholarly Catholic orders. Catholicism rested its authority on tradition to a much greater extent than did its Protestant opponents. For Catholic scholars such as Mabillon, the defense of history became really a defense of their religion. They were trying to show that historians were capable of discovering scientifically demonstrable truths. The decisive publication was Mabillon's De Re Diplomatica of 1681. A member of a rival order, the Jesuit Daniel van Papebroch, had challenged (in 1675) the authenticity of the oldest charters of two French Benedictine monasteries, Saint-Denis and Corbie. Mabillon applied his powerful critical intelligence not only to vindicating these documents but also to formulating the general rules that must be used to prove the authenticity of medieval records. He illustrated his rules by admirable examples and stated his conclusions with a candor and a common sense that convinced most readers. Mabillon's survey of the tests that must be applied by scholars covered the writing materials, the scripts (thus founding the science of medieval Latin paleography), the seals and other

devices of authentication, the official formulas, and the vocabulary used at different periods. Above all, he stressed that the authenticity of a document usually rested not just on isolated details but on consistent correctness of all its features.

Mabillon was not just a "historical scientist." He had a passionate interest in the past and a vivid historical imagination. He displayed these qualities abundantly in his last and most important work, the Annales Ordinis s. Benedicti ("Annals of the Benedictine Order," to 1066). In the Traité des études monastiques (1691; "Manual of Monastic Studies"), he defended the importance of scholarly work as the principal activity of an elite of Benedictine monks. But it would be an anachronism to regard Mabillon and his chief associates as fully comparable to modern historians. They were constrained by the limitations of their time and of their special position as monks. For example, Bernard de Montfaucon, Mabillon's most important successor, is the creator of the science of medieval Greek paleography. But he shares with most of his contemporaries a complete inability to treat the Old Testament as a historical source.

Developments in 17th-century England.

Historical and antiquarian studies developed in 17th-century England in several very distinctive ways. The political struggles and religious controversies of that period made some issues of older English history into matters of immediate practical importance. The other distinctive feature was the delay in the absorption of European continental learning, so that the great progress made in the study of feudal origins in the 16th century began to affect the thinking of English scholars only by about 1625. But there persisted also elements of continuity growing out of earlier Tudor scholarship. The interest in the Anglo-Saxon church and civilization continued to stimulate important editions of records throughout the 17th and early 18th centuries, including, especially, Sir Henry Spelman's edition of the records of church councils and Sir William Dugdale's Monasticon Anglicanum (1655-73), which is still valuable today. Another element of continuity with the Tudor period was the perennial interest of the English notables in heraldry, genealogy, and the

antiquities of their native regions. Dugdale's Antiquities of Warwickshire (1656) set a pattern and a standard for county histories.

Students of English law and institutions, lacking the stimulus that was provided for French lawyers by the diversity of legal systems and by the notable progress in the study of Roman law in that country, continued to ascribe immemorial origins to the common law of England and to approach the development of English institutions in a completely unhistorical spirit. Among the parliamentary opposition to the Stuarts, these attitudes were part of a belief in the "ancient constitution," which these sovereigns were supposed to be defying. Spelman, who was a devout Anglican and a royalist, though a moderate one, was perhaps the first major scholar to break away from this myth. Under the influence of continental publications and correspondents, he accepted that feudal tenure had been introduced into England after the Norman Conquest and that all the English institutions after 1066 must be redefined in feudal terms. But his discoveries were hidden in a dictionary of antiquarian words (Archaeologus, vol. 1, 1626; 2 vol. 1664) and made very little impact until some 50 years had elapsed. Spelman had an acute sense of historical development, and he sadly castigated his countrymen for their lack of it in their attitude to parliamentary origins: when States are departed from their original Constitution and that original by tract of time worn out of memory; the succeeding Ages viewing what is past by the present, conceive the former to have been like to that they live in. (Of Parliaments, written in about 1640, published 1698.)

His greatest contribution to English history was to grasp that parliaments had developed out of feudal assemblies convoked by the Norman kings and that the Commons were introduced into parliaments subsequently, as a result of the growing prosperity of the lesser landholders. These views first became generally accessible in the 1664 edition of Spelman's dictionary. They were adopted by Robert Brady (in 1681) and by other partisans of the Stuarts and expanded into a Royalist statement of the English past. Violently polemical though this view was, it did at least lay to rest the myth of the immemorial "ancient constitution." The Whig triumph at the Glorious Revolution of 1688, which established a doctrine

that the king ruled by parliamentary consent, led to the neglect of these discoveries for much of the 18th century. This was the common fate of much of the research of 17th-century antiquarians, who were very much ahead of their time and were writing for a limited audience. John Aubrey's pioneer description in the 1670s of the prehistoric sites of Avebury and Stonehenge had to wait two centuries for full publication. Even the best of these antiquarians, such as Spelman and Dugdale, were less critical in their handling of the original sources than Mabillon was. Higher standards were reached by a few of their successors in the early 18th century, especially by Thomas Madox, whose Formulare Anglicanum (1702) imitated Mabillon by attempting a systematic introduction to English medieval documents. But this did not save Madox from prolonged oblivion. After about 1730 this English tradition of antiquarian scholarship largely ended and remained unfashionable for most of the 18th century.

HISTORIOGRAPHY IN THE AGE OF THE ENLIGHTENMENT

The impulse given to historiography by the Italian Humanists and the religious controversialists had largely spent itself by about 1715. Men knew again how to write rationally satisfying contemporary histories, though often it needed courage to do so. Much less progress had been achieved in reconstructing the more distant past. Impressive collections of historical materials were being accumulated, but most scholars still lacked the capacity to rethink the thoughts of past generations and thus really to understand them. Mabillon could write with insight about early Benedictine history, as he possessed both sympathy with the subject and adequate technical expertise, but he was exceptional. Spelman had grasped that a particular society would be molded in a peculiar way by its institutions. He could not reconstruct and explain the gradual changes from one set of institutions to a later one, but he was aware of the problem.

Judged by the quality of its historical output, the 18th century was not, on the whole, an age of successful historians, but some of the defects of earlier historiography were beginning to be overcome. There were also losses, however,

for some of the achievements of the preceding period were in danger of being forgotten. In the leading countries of western Europe, religious controversies were becoming less important, and a massive secularization of interests took place, which affected even ecclesiastical scholars. The French Maurists continued until 1790 to publish imposing historical collections, but their choice of subjects was determined much less than in the time of Mabillon by religious priorities. The greatest Italian ecclesiastical disciple of Mabillon was Ludovico Antonio Muratori, a social reformer. In a divided country like Italy, the best way of expressing his patriotism lay in reminding Italians of the former greatness of their country. Muratori spent much of his long life on his editions of Italian medieval sources.

The nationalist motivation shown by Muratori was peculiar to Italy and also to parts of Germany, another divided country. Elsewhere in Europe there was a danger that, as men lost interest in constitutional or religious disputes that might be settled by appeals to the past, they might turn away altogether from history or at least neglect long stretches of it. This did happen to some extent in the 18th century. Some of the radical French reformers, such as Jean Le Rond d' Alembert, one of the main inspirers in the 1750s of the French Encyclopédie, wanted to jettison completely much of the past.

The Marquis de Condorcet, an early prophet of the doctrine of endless progress of mankind and a pioneer historian of European civilization, was a prominent member of a French parliamentary commission that in 1792-93 deliberately destroyed some of the royal records as comprising relics of past servitude.

During much of the 18th century it was safer and easier to publish controversial works of history than it had been in the past. The point is important, as without this greater freedom, the peculiarly radical "philosophical" historiography, so typical of that century, would have been inconceivable. In Italy such writing was still dangerous. Pietro Giannone, the author of an anticlerical history of Naples (1723), was tracked down by the Inquisition and spent 12 years in prison, where he died in 1748. Even the great Muratori, who tried to help Giannone, came into danger of having some of his works banned and had to be rescued by the personal

intervention of Pope Benedict XIV. In France, Louis XIV in 1714 imprisoned Nicolas Fréret in the Bastille for alleging (correctly) that the Franks were originally a confederacy of German tribes and not descendants of more illustrious ancestors. Under the successors of Louis, nothing quite so absurd happened again, but critics of the government or the church were often in trouble. Great Britain, Holland, Switzerland, and parts of Germany, on the other hand, provided safe oases where most things could be published. It was no accident that the most independent and historically minded group of German professors should have congregated at the University of Göttingen, founded in 1734, in the Hanoverian territory of the kings of Great Britain.

A real renewal of historiography in the 18th century could only come if fresh reasons were discovered for making it again worthwhile. Nationalism could supply one such motive; but this only became decisively influential in the 19th century. An alternative was a historiography inspired by the progress in the natural sciences and based on formulating the general rules governing the development of human societies. The chief features of this "new" historiography were a sense of the unity of all human history, including an interest in the continents outside Europe; a capacity for bold generalizations about the salient features of particular periods or societies; and a preference for topics connected with the progress of human civilization. Condorcet's historical sketch of the progress of the human mind, written in 1794, subdivided all known history into nine periods, each starting with some great invention or with geographical discoveries.

The shortcomings of this "rationalistic" historiography have been rehearsed often enough. For many of its writers it was primarily a weapon of propaganda against their enemies in church and state. Their redeeming virtue was the fearlessly critical attitude to all existing authorities, however august or sacred. The vast scale of their generalizations often precluded any detailed research. This was particularly true of the attempts to write histories of civilization, as the existing collections of printed materials did not cater for such interests, while systematic research in archives was seldom possible in the 18th century. In preparing his pioneer essay on the history

of civilization, covering the millennium from the Carolingians to Louis XIV (Essai sur les moeurs et l'esprit des nations, 1745-53), the French author Voltaire had to collect bits and pieces from most diverse sources.

One of the most valuable achievements of the thinkers of the 18th century was their capacity to study particular societies as coherent units and to formulate the theory that the various aspects of each society's life were closely interrelated. This was not an entirely novel idea, but it first became commonly accepted during this period. Nor were all its adherents anticlericals. Giambattista Vico, a Neapolitan Catholic, was ahead of his contemporaries in his particularly subtle sense of the complex influences by which one phase of society gives place to another. In his reconstruction of these transitions during the early stages of Roman history, he makes no clear lines between periods. His countryman Giannone explains in his autobiography that he had studied Roman law not for its own sake but in order to understand the changes in the society of the Roman Empire. The French philosopher Montesquieu, who owed much to Giannone, was not really a historian, but he displays an acute sense of historical realities. His De l'esprit des lois (1748; The Spirit of Laws), more than any other book, accustomed his contemporaries to ponder the complex factors that shaped each society. It inspired Gibbon's definition of the kind of history he wanted to write. It was to be a "history related to and explained by the social institutions in which it is contained."

This ideal was realized in Gibbon's History of the Decline and Fall of the Roman Empire (1776-88), one of the masterpieces of "philosophical" historiography. Gibbon was preoccupied above all with the problem of human progress. The belief that continuous progress was possible for mankind had been publicly formulated in the mid-18th century by Anne-Robert-Jacques Turgot in France and by Adam Smith in Scotland, independently, it seems, of each other. Gibbon had read works and known scholars influenced by both these thinkers. A belief in continuous progress would confer a new purposefulness on the study of the entire course of human history and could justify a lengthy account of what otherwise might have seemed very obscure stretches of the past. Such a justification was to inspire most

of the historiography of the 19th century. But the problem of progress had a special urgency for Gibbon's generation, which worried at the thought that their own enlightened civilization might also subsequently collapse. By unravelling the causes of the decline of the Roman Empire, Gibbon was determined to show that the Europe of his own day had attained a much superior degree of development and was immune from the fate of the ancient world.

In the 18th century, historiography was still only very rarely connected with the universities; and thus, except in such isolated places as Göttingen in Germany, no continuous schools of history could develop. Some of the most important achievements of the 18th-century historians meant much less to their contemporaries than to their successors in the 19th century. Gibbon was a pioneer in utilizing in a "rationalist" history the vast materials accumulated by generations of erudite antiquarians, but he had no immediate followers. The German archaeologist Johann Joachim Winckelmann tried to revive the true understanding of Greek sculpture and to make the history of art into something more than just the biographies of artists, but his work bore little fruit until the next century. The saddest fate was that of Vico's work. He was hardly ever read before the 19th century, when he at last influenced Barthold Georg Niebuhr and the rest of the German historical school, while Jules Michelet's rediscovery of Vico in 1824 started a new era in French writing on the Middle Ages.

HISTORIOGRAPHY IN THE 19TH AND 20TH CENTURIES

Growth of specialization.

From the early 19th century, historiography began to develop in a radically different way. The decisive changes occurred among the German historians, largely through a reaction to the French Revolution and to a temporary subjugation of their country by Napoleon. Organized teaching of history in schools and universities became a matter of national importance, first in Prussia and then in other parts of Germany. As universal education spread to most European countries in the course of the 19th century, history was accepted everywhere as a necessary

subject in schools. For the first time the bulk of historical writing came to be done by professional historians, for whom it became a condition of securing academic appointments or of consolidating their standings as university teachers. Historiography eventually became a continuously cooperative venture, where the achievements of past historians could be used systematically by their successors. But the growth of specialization and the bewildering number of types of works that came to be published constituted a new danger. In the past, important discoveries were frequently lost through lack of interest. But, by the second half of the 20th century, discoveries were in danger of being simply overlooked amid the flood of publications.

Another great change lay in the growth of intellectual freedom. Free expression of independent or unorthodox ideas had become dangerous during the French Revolution and under Napoleon, both in the territories controlled by the French and, by way of frightened reaction, in the lands of their unconquered opponents. After 1815 conditions for freer historiography improved gradually in much of Europe. Charles Darwin's Origin of Species (1859), which put forth a theory of evolution at first unacceptable to church authorities, probably could not have been published with the same impunity any earlier.

One feature of the growing tolerance of governments toward historiography was the gradual creation of public archives, such as the British Public Record Office in London, created in 1838, and the freer opening of the collections already in existence. Even the papacy accepted these changes, and Pope Leo XIII opened up the papal archive in 1883 as part of a deliberate new policy of encouraging historical study of Catholicism. For the first time historiography came to be based largely on unpublished records, and scholars were tempted into excessive reliance on original documents while unduly neglecting the older types of narrative sources.

In the 20th century some grievous threats to the persistence of free scholarship recurred, and historiography suffered with other branches of humane studies. The establishment of a Communist regime in Russia led, at first, to the rejection of most pre-1917 history as a fit subject for schools and universities. This decision

was reversed in the 1930s, and from 1945 Communist countries were encouraging a form of historiography especially concerned with economic history and the class struggles of the past. There was also an enthusiastic interest in the material remains of past ages, leading to an impressive development of archaeology, particularly in Poland. The rise of dictatorships in Italy and Germany had disastrous effects on historiography in those countries, and recovery after World War II was only gradual.

Judged merely by the number of "practicing" historians and of their publications, historiography seemed in a very flourishing state in the 1970s. Its European traditions had spread to all the other continents and were largely accepted in all non-Communist countries.

The Introduction aux études historiques (Introduction to the Study of History) of Charles V. Langlois and Charles Seignobos (1898), supplemented by critical comments of another outstanding French historian, Ferdinand Lot (in Le Moyen Age, 1898), provides an excellent starting point for the discussion of modern historical methods. History is an autonomous branch of learning, and some of its methods may be unique. Historians should not try to formulate general laws; their branch of learning merely "aims at explaining reality." Langlois and Seignobos particularly stress that history is not a science of observation but a science of reasoning how to extract from imperfect documentary or narrative records some glimpses of what actually happened.

The historian's task.

A historian has to subject his sources to a whole series of preliminary investigations. First comes "external criticism," aimed at determining whether the sources are appropriate and adequate for the particular task in hand. The provenance, date, and authenticity of each source must be established by using the techniques of diplomatic, the detailed study and assessment of documents, and of paleography, the study of ancient handwriting, and of other auxiliary sciences that were elaborated after the 17th century. In France a special institution for teaching some of these techniques, the École des Chartes, was created in 1821. The first

specialized seminar for instruction in these subjects was established in 1854 at Vienna by Theodor von Sickel, one of the greatest medievalists of the 19th century, and it was gradually imitated by leading German universities.

One of the most important critical refinements introduced in the course of the 19th century was the improved handling of narrative sources brought about by seeking to discover the literary sources that lay behind them. Leopold von Ranke, one of the foremost German historians, who began his career as a teacher of classics, was gradually attracted to history through a desire to understand better the sources of the Greek and Latin authors whom he was expounding. In the later decades of the 19th century, such a quest became a normal feature of historical scholarship.

Once a historian has decided, through the application of "external criticism," on the sources that are relevant to his purpose, he must next, by "internal criticism," make sure that he fully understands what he has selected. German classical philologists were the first to bring these latter investigations to a high degree of perfection. Karl Lachmann, an editor of the Latin poets, is justly regarded as the creator of modern textual criticism in its most rigorous forms, and historians gradually adopted similar methods. The language of the sources must be understood, corruptions in the text must be eliminated, and the historian must, as accurately as possible, penetrate the minds of the authors with whom he is dealing.

All these critical operations on the sources are merely preliminaries, and the work of the historian proper only starts when he attempts a synthesis of his materials. F. Lot stresses that in this qualities other than the erudite skills come into play. There must be sympathy with the subjects under study, for without it there can be no imaginative insight into the past. Ideally, a historian must display capacities akin to those of a poet or an artist.

The Social Sciences

Introduction

The social sciences, which deal with human behaviour in its social and cultural aspects, include the following disciplines: cultural (or social) anthropology, sociology, social psychology, political science, and economics. Also frequently included are social and economic geography and those areas of education that deal with the social contexts of learning and the relation of the school to the social order.

History is regarded by many as a social science, and certain areas of historical study are almost indistinguishable from work done in the social sciences. Most historians, however, still consider history as one of the humanities. It is generally best, in any case, to consider history as marginal to the humanities and social sciences, since its insights and techniques pervade both.

The study of comparative law may also be regarded as a part of the social sciences, although it is ordinarily pursued in schools of law rather than in departments or schools containing most of the other social sciences.

Since the 1950s the term behavioral sciences has often been applied to the disciplines designated as the social sciences. Those who favour this term do so in part because these disciplines are thus brought closer to some of the sciences, such as physical anthropology and physiological psychology, which also deal with human behaviour. Whether the term behavioral sciences will in time supplant "social sciences" or whether it will, as neologisms so often have before, fade away is impossible to say. For the purposes of this article, the two terms may be considered synonymous.

This article is concerned with the social sciences as vital elements in the aftermath of the two great revolutions, the political and industrial, which opened the 19th century, with the pattern the social sciences assumed in that century, and their development in the 20th century.

History of the social sciences

Although, strictly speaking, the social sciences do not precede the 19th century - that is, as distinct and recognized disciplines of thought - one must go back farther in time for the origins of some of their fundamental ideas and objectives. In the largest sense, the origins go all the way back to the ancient Greeks and their rationalist inquiries into the nature of man, state, and morality. The heritage of both Greece and Rome is a powerful one in the history of social thought as it is in so many other areas of Western society. Very probably, apart from the initial Greek determination to study all things in the spirit of dispassionate and rational inquiry, there would be no social sciences today. True, there have been long periods of time, as during the Western Middle Ages, when the Greek rationalist temper was lacking.

But the recovery of this temper, through texts of the great classical philosophers, is the very essence of the Renaissance and the Age of Reason in modern European history. With the Age of Reason, in the 17th and 18th centuries, one may begin.

HERITAGE OF THE MIDDLE AGES AND THE RENAISSANCE

Effects of theology.

The same impulses that led men in that age to explore the earth, the stellar regions, and the nature of matter led them also to explore the institutions around them: state, economy, religion, morality; above all, the nature of man himself. It was the fragmentation of medieval philosophy and theory, and, with this, the shattering of the medieval world view that had lain deep in thought until about the 16th century, that was the immediate basis of the rise of the several strands of specialized thought that were to become in time the social sciences.

Medieval theology, especially as it appears in St. Thomas Aquinas' Summa theologiae, contained and fashioned syntheses from ideas about man and society - ideas indeed that may be seen to be political, social, economic, anthropological, and geographical in their substance. But it was partly this close relation between medieval theology and ideas of the social sciences that accounts for the longer time it took these ideas - by comparison with the ideas of the physical sciences - to

achieve what one would today call scientific character. From the time of the great Roger Bacon in the 13th century, there were at least some rudiments of physical science that were largely independent of medieval theology and philosophy. Historians of physical science have no difficulty in tracing the continuation of this experimental tradition, primitive and irregular though it was by later standards, throughout the Middle Ages. Side by side with the kinds of experiment made notable by Roger Bacon were impressive changes in technology through the medieval period and then, in striking degree, in the Renaissance. Efforts to improve agricultural productivity; the rising utilization of gunpowder, with consequent development of guns and the problems that they presented in ballistics; growing trade, leading to increased use of ships and improvements in the arts of navigation, including use of telescopes; and the whole range of such mechanical arts in the Middle Ages and Renaissance as architecture, engineering, optics, and the construction of watches and clocks - all of this put a high premium on a pragmatic and operational understanding of at least the simpler principles of mechanics, physics, astronomy, and, in time, chemistry.

In short, by the time of Copernicus and Galileo in the 16th century, a fairly broad substratum of physical science existed, largely empirical but not without theoretical implications on which the edifice ofmodern physical science could be built. It is notable that the empirical foundations of physiology were being established in the studies of the human body being conducted in medieval schools of medicine and, as the career of Leonardo da Vinci so resplendently illustrates, among artists of the Renaissance, whose interest in accuracy and detail of painting and sculpture led to their careful studies of human anatomy.

Very different was the beginning of the social sciences. In the first place, the church, throughout the Middle Ages and even into the Renaissance and Reformation, was much more attentive to what scholars wrote and thought about man's mind and his behaviour in society than it was toward what was being studied and written in the physical sciences. From the church's point of view, while it might be important to see to it that thought on the physical world corresponded as

far as possible to what Scripture said - witnessed, for example, in the famous questioning of Galileo - it was far more important that such correspondence exist in matters affecting the nature of man, his mind, spirit, and soul. Nearly all the subjects and questions that would form the bases of the social sciences in later centuries were tightly woven into the fabric of medieval scholasticism, and it was not easy for even the boldest minds to break this fabric.

Effects of the classics and of Cartesianism.

Then, when the hold of scholasticism did begin to wane, two fresh influences, equally powerful, came on the scene to prevent anything comparable to the pragmatic and empirical foundations of the physical sciences from forming in the study of man and society. The first was the immense appeal of the Greek classics during the Renaissance, especially those of the philosophers Plato and Aristotle. A great deal of social thought during the Renaissance was little more than gloss or commentary on the Greek classics. One sees this throughout the 15th and 16th centuries.

Second, in the 17th century appeared the powerful influence of the philosopher René Descartes. Cartesianism, as his philosophy was called, declared that the proper approach to understanding of the world, including man and society, was through a few simple, fundamental ideas of reality and, then, rigorous, almost geometrical deduction of more complex ideas and eventually of large, encompassing theories, from these simple ideas, all of which, Descartes insisted, were the stock of common sense - the mind that is common to all human beings at birth. It would be hard to exaggerate the impact of Cartesianism on social and political and moral thought during the century and a half following publication of his Discourse on Method and his Meditations. Through the Age of Reason and down through the Enlightenment in the later 18th century, the spell of Cartesianism was cast on nearly all those who were concerned with the problems of the nature of man and society.

Both of these great influences, reverence for the classics and fascination with the geometrical-deductive procedures advocated by Descartes must be seen from today's vantage point as among the major influences retarding the development of a science of society comparable to the science of the physical world. It is not as though data were not available in the 17th and 18th centuries. The emergence of the national state carried with it evergrowing bureaucracies concerned with gathering information, chiefly for taxation, census, and trade purposes, which might have been employed in much the same way that physical scientists employed their data. The voluminous and widely published accounts of the great voyages that had begun in the 15th century, the records of soldiers, explorers, and missionaries who perforce had been brought into often long and close contact with primitive and other non-Western peoples, provided still another great reservoir of data, all of which might have been utilized in scientific ways as such data were to be utilized a century or two later in the social sciences. Such, however, was the continuing spell cast by the texts of the classics and by the strictly rationalistic, overwhelmingly deductive procedures of the Cartesians that, down until the beginning of the 19th century, these and other empirical materials were used, if at all, solely for illustrative purposes in the writings of the social philosophers.

HERITAGE OF THE ENLIGHTENMENT

There is also the fact that, especially in the 18th century, reform and even revolution were often in the air. The purpose of a great many social philosophers was by no means restricted to philosophic, much less scientific, understanding of man and society. The dead hand of the Middle Ages seemed to many vigorous minds in western Europe the principal force to be combatted, through critical reason, enlightenment, and, where necessary, major reform or revolution. One may properly account a great deal of this new spirit to the rise of humanitarianism in modern Europe and in other parts of the world and to the spread of literacy, the rise in the standard of living, and the recognition that poverty and oppression need not be the fate of the masses. The fact remains, however, that social reform and social

science have different organizing principles, and the very fact that for a long time, down indeed through a good part of the 19th century, social reform and social science were regarded as pretty much the same thing could not have helped but retard the development of the latter.

Nevertheless, it would be wrong to discount the significant contributions to the social sciences that were made during the 17th and 18th centuries. The first and greatest of these was the spreading ideal of a science of society, an ideal fully as widespread by the 18th century as the ideal of a physical science. Second was the rising awareness of the multiplicity and variety of human experience in the world. Ethnocentrism and parochialism, as states of mind, were more and more difficult for educated people to maintain given the immense amount of information about - or, more important, interest in - non-Western peoples, the results of trade and exploration. Third was the spreading sense of the social or cultural character of human behaviour in society - that is, its purely historical or conventional, rather than biological, basis. A science of society, in short, was no mere appendage of biology but was instead a distinct discipline, or set of disciplines, with its own distinctive subject matter.

To these may be added two other very important contributions of the 17th and 18th centuries, each of great theoretical importance. The first was the idea of structure. First seen in the writings of such philosophers as Hobbes, Locke, and Rousseau with reference to the political structure of the state, it had spread by the mid-18th century to highlight the economic writings of the Physiocrats and Adam Smith. The idea of structure can also be seen in certain works relating to man's psychology and, at opposite reach, to the whole of civil society. The ideas of structure that were borrowed from both the physical and biological sciences were fundamental to the conceptions of political, economic, and social structure that took shape in the 17th and 18th centuries. And these conceptions of structure have in many instances, subject only to minor changes, come down to 20th-century social science.

The second major theoretical idea was that of developmental change. Its ultimate roots in Western thought, like those indeed of the whole idea of structure, go back to the Greeks, if not earlier. But it is in the 18th century, above all others, that the philosophy of developmentalism took shape, forming a preview, so to speak, of the social evolutionism of the next century. What was said by such writers as Condorcet, Rousseau, and Adam Smith was that the present is an outgrowth of the past, the result of a long line of development in time, and, furthermore, a line of development that has been caused, not by God or fortuitous factors, but by conditions and causes immanent in human society. Despite a fairly widespread belief that the idea of social development is a product of prior discovery of biological evolution, the facts are the reverse. Well before any clear idea of genetic speciation existed in European biology, there was a very clear idea of what might be called social speciation - that is, the emergence of one institution from another in time and of the whole differentiation of function and structure that goes with this emergence.

As has been suggested, these and other seminal ideas were contained for the most part in writings, the primary function of which was attack on the existing order of government and society in western Europe. Another way of putting the matter is to say that they were clear and acknowledged parts of political and social idealism - using that word in its largest sense. Hobbes, Locke, Rousseau, Montesquieu, Adam Smith, and other major philosophers had as vivid and energizing sense of the ideal - ideal state, ideal economy, ideal civil society - as any earlier utopian writer. These men were, without exception, committed to visions of the good or ideal society. Their interest in the "natural" - that is, natural morality, religion, economy, or education, in contrast to the merely conventional and historically derived - sprang as much from the desire to hold a glass up to a surrounding society that they disliked as from any dispassionate urge simply to find out what man and society are made of. The fact remains, however, that the ideas that were to prove decisive in the 19th century, so far as the social sciences were concerned, arose during the two centuries preceding.

THE 19TH CENTURY

The fundamental ideas, themes, and problems of the social sciences in the 19th century are best understood as responses to the problem of order that was created in men's minds by the weakening of the old order, or European society, under the twin blows of the French Revolution and the Industrial Revolution. The breakup of the old order - an order that had rested on kinship, land, social class, religion, local community, and monarchy - set free, as it were, the complex elements of status, authority, and wealth that had been for so long consolidated. In the same way that the history of 19th-century politics, industry, and trade is basically about the practical efforts of human beings to reconsolidate these elements, so the history of 19th-century social thought is about theoretical efforts to reconsolidate them - that is, to give them new contexts of meaning.

In terms of the immediacy and sheer massiveness of impact on human thought and values, it would be difficult to find revolutions of comparable magnitude in human history. The political, social, and cultural changes that began in France and England at the very end of the 18th century spread almost immediately through Europe and the Americas in the 19th century and then on to Asia, Africa, and Oceania in the 20th. The effects of the two revolutions, the one overwhelmingly democratic in thrust, the other industrial-capitalist, have been to undermine, shake, or topple institutions that had endured for centuries, even millennia, and with them systems of authority, status, belief, and community.

It is easy today to deprecate the suddenness, the cataclysmic nature, the overall revolutionary effect of these two changes and to seek to subordinate results to longer, deeper tendencies of more gradual change in western Europe. But as many recent historians have pointed out, there was to be seen, and seen by a great many sensitive minds of that day, a dramatic and convulsive quality to the changes that cannot properly be subsumed to the slower processes of continuous evolutionary change. What is crucial, in any event, from the point of view of the history of the social thought of the period, is how the changes were actually envisaged at the

time. By a large number of social philosophers and social scientists, in all spheres, those changes were regarded as nothing less than of earthquake intensity.

The coining or redefining of words is an excellent indication of men's perceptions of change in a given historical period. A large number of words taken for granted today came into being in the period marked by the final decade or two of the 18th century and the first quarter of the 19th. Among these are: industry, industrialist, democracy, class, middle class, ideology, intellectual, rationalism, humanitarian, atomistic, masses, commercialism, proletariat, collectivism, equalitarian, liberal, conservative, scientist, utilitarian, bureaucracy, capitalism, and crisis. Some of these words were invented; others reflect new and very different meanings given to old ones. All alike bear witness to the transformed character of the European social landscape as this landscape loomed up to the leading minds of the age. And all these words bear witness too to the emergence of new social philosophies and, most pertinent to the subject of this article, the social sciences as they are known today.

Major themes resulting from democratic and industrial change.

It is illuminating to mention a few of the major themes in social thought in the 19th century that were almost the direct results of the democratic and industrial revolutions. It should be borne in mind that these themes are to be seen in the philosophical and literary writing of the age as well as in social thought.

First, there was the great increase in population. Between 1750 and 1850 the population of Europe went from 140,000,000 to 266,000,000; in the world from 728,000,000 to well over 1,000,000,000. It was an English clergyman-economist, Thomas Malthus, who, in his famous Essay on Population, first marked the enormous significance to human welfare of this increase. With the diminution of historic checks on population growth, chiefly those of high mortality rates - a diminution that was, as Malthus realized, one of the rewards of technical progress - there were no easily foreseeable limits to growth of population. And such growth, he stressed, could only upset the traditional balance between population, which Malthus described as growing at geometrical rate, and food supply, which he

declared could grow only at arithmetical rate. Not all social scientists in the century took the pessimistic view of the matter that Malthus did but few if any were indifferent to the impact of explosive increase in population on economy, government, and society.

Second, there was the condition of labour. It may be possible to see this condition in the early 19th century as in fact better than the condition of the rural masses at earlier times. But the important point is that to a large number of writers in the 19th century it seemed worse and was defined as worse. The wrenching of large numbers of people from the older and protective contexts of village, guild, parish, and family, and their massing in the new centres of industry, forming slums, living in common squalor and wretchedness, their wages generally behind cost of living, their families growing larger, their standard of living becoming lower, as it seemed - all of this is a frequent theme in the social thought of the century. Economics indeed became known as the "dismal science," because economists, from David Ricardo to Karl Marx, could see little likelihood of the condition of labour improving under capitalism.

Third, there was the transformation of property. Not only was more and more property to be seen as industrial - manifest in the factories, business houses, and workshops of the period - but also the very nature of property was changing. Whereas for most of the history of mankind property had been "hard," visible only in concrete possessions - land and money - now the more intangible kinds of property such as shares of stock, negotiable equities of all kinds, and bonds were assuming ever greater influence in the economy. This led, as was early realized, to the dominance of financial interests, to speculation, and to a general widening of the gulf between the propertied and the masses. The change in the character of property made easier the concentration of property, the accumulation of immense wealth in the hands of a relative few, and, not least, the possibility of economic domination of politics and culture. It should not be thought that only socialists saw property in this light. From Edmund Burke through Auguste Comte, Frédéric Le

Play, and John Stuart Mill down to Karl Marx, Max Weber, and Émile Durkheim, one finds conservatives and liberals looking at the impact of this change in analogous ways.

Fourth, there was urbanization - the sudden increase in the number of towns and cities in western Europe and the increase in number of persons living in the historic towns and cities. Whereas in earlier centuries, the city had been regarded almost uniformly as a setting of civilization, culture, and freedom of mind, now one found more and more writers aware of the other side of cities: the atomization of human relationships, broken families, the sense of the mass, of anonymity, alienation, and disrupted values. Sociology particularly among the social sciences turned its attention to the problems of urbanization. The contrast between the more organic type of community found in rural areas and the more mechanical and individualistic society of the cities is a basic contrast in sociology, one that was given much attention by such pioneers in Europe as the French sociologists Frédéric Le Play and Émile Durkheim; the German sociologists Ferdinand Tönnies, Georg Simmel, and Max Weber; the Belgian statistician Adolphe Quetelet; and, in America, by the sociologists Charles H. Cooley and Robert E. Park.

Fifth, there was technology. With the spread of mechanization, first in the factories, then in agriculture, social thinkers could see possibilities of a rupture of the historic relation between man and nature, between man and man, even between man and God. To thinkers as politically different as Thomas Carlyle and Karl Marx, technology seemed to lead to dehumanization of the worker and to exercise of a new kind of tyranny over human life. Marx, though, far from despising technology, thought the advent of socialism would counteract all this. Alexis de Tocqueville declared that technology, and especially technical specialization of work, was more degrading to man's mind and spirit than even political tyranny. It was thus in the 19th century that the opposition to technology on moral, psychological, and aesthetic grounds first made its appearance in Western thought.

Sixth, there was the factory system. The importance of this to 19th-century thought has been intimated above. Suffice it to add that along with urbanization and spreading mechanization, the system of work whereby masses of workers left home and family to work long hours in the factories became a major theme of social thought as well as of social reform.

Seventh, and finally, mention is to be made of the development of political masses - that is, the slow but inexorable widening of franchise and electorate through which ever larger numbers of persons became aware of themselves as voters and participants in the political process. This too is a major theme in social thought, to be seen most luminously perhaps in Tocqueville's Democracy in America, a classic written in the 1830s that took not merely America but democracy everywhere as its subject. Tocqueville saw the rise of the political masses, more especially the immense power that could be wielded by the masses, as the single greatest threat to individual freedom and cultural diversity in the ages ahead.

These, then, are the principal themes in the 19th-century writing that may be seen as direct results of the two great revolutions. As themes, they are to be found not only in the social sciences but, as noted above, in a great deal of the philosophical and literary writing of the century. In their respective ways, the philosophers Hegel, Coleridge, and Emerson were as struck by the consequences of the revolutions as were any social scientists. So too were such novelists as Balzac and Dickens.

New ideologies.

One other point must be emphasized about these themes. They became, almost immediately in the 19th century, the bases of new ideologies. How men reacted to the currents of democracy and industrialism stamped them conservative, liberal, or radical. On the whole, with rarest exceptions, liberals welcomed the two revolutions, seeing in their forces opportunity for freedom and welfare never before known to mankind. The liberal view of society was overwhelmingly democratic, capitalist, industrial, and, of course, individualistic. The case is somewhat different with conservatism and radicalism in the century.

Conservatives, beginning with Edmund Burke, continuing through Hegel and Matthew Arnold down to such minds as John Ruskin later in the century, disliked both democracy and industrialism, preferring the kind of tradition, authority, and civility that had been, in their minds, displaced by the two revolutions. Theirs was a retrospective view, but it was a nonetheless influential one, affecting a number of the central social scientists of the century, among them Auguste Comte and Tocqueville and later Max Weber and Émile Durkheim. The radicals accepted democracy but only in terms of its extension to all areas of society and its eventual annihilation of any form of authority that did not spring directly from the people as a whole. And although the radicals, for the most part, accepted the phenomenon of industrialism, especially technology, they were uniformly antagonistic to capitalism.

These ideological consequences of the two revolutions proved extremely important to the social sciences, for it would be difficult to identify a social scientist in the century - as it would a philosopher or a humanist - who was not, in some degree at least, caught up in ideological currents. In referring to such minds as Saint-Simon, Comte, Le Play among sociologists, to Ricardo, the Frenchman Jean-Baptiste Say, and Marx among economists, to Jeremy Bentham and John Austin among political scientists, even to anthropologists like the Englishman Edward B. Tylor and the American Lewis Henry Morgan, one has before one men who were engaged not merely in the study of society but also in often strongly partisan ideology. Some were liberals, some conservatives, others radicals. All drew from the currents of ideology that had been generated by the two great revolutions.

New intellectual and philosophical tendencies.

It is important also to identify three other powerful tendencies of thought that influenced all of the social sciences. The first is a positivism that was not merely an appeal to science but almost reverence for science; the second, humanitarianism; the third, the philosophy of evolution.

The Positivist appeal of science was to be seen everywhere. The rise of the ideal of science in the Age of Reason was noted above. The 19th century saw the virtual

institutionalization of this ideal - possibly even canonization. The great aim was that of dealing with moral values, institutions, and all social phenomena through the same fundamental methods that could be seen so luminously in such areas as physics and biology. Prior to the 19th century, no very clear distinction had been made between philosophy and science, and the term philosophy was even preferred by those working directly with physical materials, seeking laws and principles in the fashion of a Newton or Harvey - that is, by persons whom one would now call scientists.

In the 19th century, in contrast, the distinction between philosophy and science became an overwhelming one. Virtually every area of man's thought and behaviour was thought by a rising number of persons to be amenable to scientific investigation in precisely the same degree that physical data were. More than anyone else, it was Comte who heralded the idea of the scientific treatment of social behaviour.

His Cours de philosophie positive, published in six volumes between 1830 and 1842, sought to demonstrate irrefutably not merely the possibility but the inevitability of a science of man, one for which Comte coined the word "sociology" and that would do for man the social being exactly what biology had already done for man the biological animal. But Comte was far from alone. There were many in the century to join in his celebration of science for the study of society.

Humanitarianism, though a very distinguishable current of thought in the century, was closely related to the idea of a science of society. For the ultimate purpose of social science was thought by almost everyone to be the welfare of society, the improvement of its moral and social condition. Humanitarianism, strictly defined, is the institutionalization of compassion; it is the extension of welfare and succor from the limited areas in which these had historically been found, chiefly family and village, to society at large. One of the most notable and also distinctive aspects of the 19th century was the constantly rising number of persons, almost wholly from the middle class, who worked directly for the betterment of society. In the

many projects and proposals for relief of the destitute, improvement of slums, amelioration of the plight of the insane, the indigent, and imprisoned, and other afflicted minorities could be seen the spirit of humanitarianism at work. All kinds of associations were formed, including temperance associations, groups and societies for the abolition of slavery and of poverty and for the improvement of literacy, among other objectives. Nothing like the 19th-century spirit of humanitarianism had ever been seen before in western Europe - not even in France during the Enlightenment, where interest in mankind's salvation tended to be more intellectual than humanitarian in the strict sense. Humanitarianism and social science were reciprocally related in their purposes. All that helped the cause of the one could be seen as helpful to the other.

The third of the intellectual influences is that of evolution. It affected every one of the social sciences, each of which was as much concerned with the development of things as with their structures. An interest in development was to be found in the 18th century, as noted earlier. But this interest was small and specialized compared with 19th-century theories of social evolution. The impact of Charles Darwin's Origin of Species, published in 1859, was of course great and further enhanced the appeal of the evolutionary view of things. But it is very important to recognize that ideas of social evolution had their own origins and contexts. The evolutionary works of such social scientists as Comte, Herbert Spencer, and Marx had been completed, or well begun, before publication of Darwin's work. The important point, in any event, is that the idea or the philosophy of evolution was in the air throughout the century, as profoundly contributory to the establishment of sociology as a systematic discipline in the 1830s as to such fields as geology, astronomy, and biology. Evolution was as permeative an idea as the Trinity had been in medieval Europe.

Development of the separate disciplines.

Among the disciplines that formed the social sciences, two contrary, for a time equally powerful, tendencies at first dominated them. The first was the drive toward unification, toward a single, master social science, whatever it might be

called. The second tendency was toward specialization of the individual social sciences. If, clearly, it is the second that has triumphed, with the results to be seen in the disparate, sometimes jealous, highly specialized disciplines seen today, the first was not without great importance and must also be examined.

What emerges from the critical rationalism of the 18th century is not, in the first instance, a conception of need for a plurality of social sciences, but rather for a single science of society that would take its place in the hierarchy of the sciences that included the fields of astronomy, physics, chemistry, and biology. When, in the 1820s, Comte wrote calling for a new science, one with man the social animal as the subject, he assuredly had but a single, encompassing science of society in mind - not a congeries of disciplines, each concerned with some single aspect of man's behaviour in society. The same was true of Bentham, Marx, and Spencer. All these minds, and there were many others to join them, saw the study of society as a unified enterprise. They would have scoffed, and on occasion did, at any notion of a separate economics, political science, sociology, and so on. Society is an indivisible thing, they would have argued; so, too, must be the study of society.

It was, however, the opposite tendency of specialization or differentiation that won out. No matter how the century began, or what were the dreams of a Comte, Spencer, or Marx, when the 19th century ended, not one but several distinct, competitive social sciences were to be found. Aiding this process was the development of the colleges and universities. With hindsight it might be said that the cause of universities in the future would have been strengthened, as would the cause of the social sciences, had there come into existence, successfully, a single curriculum, undifferentiated by field, for the study of society. What in fact happened, however, was the opposite. The growing desire for an elective system, for a substantial number of academic specializations, and for differentiation of academic degrees, contributed strongly to the differentiation of the social sciences. This was first and most strongly to be seen in Germany, where, from about 1815 on, all scholarship and science were based in the universities and where competition for status among the several disciplines was keen. But by the end of

the century the same phenomenon of specialization was to be found in the United States (where admiration for the German system was very great in academic circles) and, in somewhat less degree, in France and England. Admittedly, the differentiation of the social sciences in the 19th century was but one aspect of a larger process that was to be seen as vividly in the physical sciences and the humanities. No major field escaped the lure of specialization of investigation, and clearly, a great deal of the sheer bulk of learning that passed from the 19th to the 20th century was the direct consequence of this specialization.

Economics.

It was economics that first attained the status of a single and separate science, in ideal at least, among the social sciences. That autonomy and self-regulation that the Physiocrats and Adam Smith had found, or thought they had found, in the processes of wealth, in the operation of prices, rents, interest, and wages during the 18th century became the basis of a separate and distinctive economics - or, as it was often called, "political economy" - in the 19th. Hence the emphasis upon what came to be widely called laissez-faire. If, as it was argued, the processes of wealth operate naturally in terms of their own built-in mechanisms, then not only should these be studied separately but they should, in any wise polity, be left alone by government and society. This was, in general, the overriding emphasis of such thinkers as David Ricardo, John Stuart Mill, and Nassau William Senior in England, of Frédéric Bastiat and Jean-Baptiste Say in France, and, somewhat later, the Austrian school of Carl Menger. This emphasis is today called "classical" in economics, and it is even now, though with substantial modifications, a strong position in the field.

There were almost from the beginning, however, economists who diverged sharply from this laissez-faire, classical view. In Germany especially there were the so-called historical economists. They proceeded less from the discipline of historiography than from the presuppositions of social evolution, referred to above. Such men as Wilhelm Roscher and Karl Knies in Germany tended to dismiss the

assumptions of timelessness and universality regarding economic behaviour that were almost axiomatic among the followers of Adam Smith, and they strongly insisted upon the developmental character of capitalism, evolving in a long series of stages from other types of economy.

Also prominent throughout the century were those who came to be called the Socialists. They too repudiated any notion of timelessness and universality in capitalism and its elements of private property, competition, and profit. Not only was this system but a passing stage of economic developments; it could be - and, as Marx was to emphasize, would be - shortly supplanted by a more humane and also realistic economic system based upon cooperation, the people's ownership of the means of production, and planning that would eradicate the vices of competition and conflict.

Political science.

Rivalling economics as a discipline during the century was political science. The line of systematic interest in the state that had begun in modern Europe with Machiavelli, Hobbes, Locke, and Rousseau, among others, widened and lengthened in the 19th century, the consequence of the two revolutions. If the Industrial Revolution seemed to supply all the problems frustrating the existence of a stable and humane society, the political-democratic revolution could be seen as containing many of the answers to these problems. It was the democratic revolution, especially in France, that created the vision of a political government responsible for all aspects of human society and, most important, possessed the power to wield this responsibility. This power, known as sovereignty, could be seen as holding the same relation to political science in the 19th century that capital held to economics. To a very large number of political scientists, the aim of the discipline was essentially that of analyzing the varied properties of sovereignty. There was a strong tendency on the part of such political scientists as Bentham, Austin, and Mill in England and Francis Lieber and Woodrow Wilson in the

United States to see the state and its claimed sovereignty over human lives in much the same terms in which classical economists saw capitalism.

Among political scientists there was the same historical-evolutionary dissent from this view, however, that existed in economics. Such writers as Sir Henry Maine in England, Numa Fustel de Coulanges in France, and Otto von Gierke in Germany declared that state and sovereignty were not timeless and universal nor the results of some "social contract" envisaged by such philosophers as Locke and Rousseau but, rather, structures formed slowly through developmental or historical processes. Hence the strong interest, especially in the late 19th century, in the origins of political institutions in kinship, village, and caste, and in the successive stages of development that have characterized these institutions. In political science, as in economics, in short, the classical analytical approach was strongly rivalled by the evolutionary. Both approaches go back to the 18th century in their fundamental elements, but what is seen in the 19th century is the greater systematization and the much wider range of data employed.

Cultural anthropology.

In the 19th century, anthropology also attained clear identity as a discipline. Strictly defined as "the science of man," it could be seen as superseding other specialized disciplines such as economics and political science. In practice and from the beginning, however, anthropology concerned itself overwhelmingly with primitive man. On the one hand was physical anthropology, concerned chiefly with the evolution of man as a biological species, with the successive forms and protoforms of the species, and with genetic systems such as stocks and races in the world. On the other hand was social and

cultural anthropology: here the interest was in the full range of man's institutions but confined to those found in fact among existing preliterate or "primitive" peoples in Africa, Oceania, Asia, and the Americas.

Above all other concepts, "culture" was the central element of this great area of anthropology, or ethnology, as it was often called to distinguish it from physical

anthropology. Culture, as a concept, called attention to the nonbiological, nonracial, noninstinctual basis of the greater part of what one calls civilization: its values, techniques, ideas in all spheres. Culture, as defined in Tylor's landmark work of 1871, Primitive Culture, is the part of man's behaviour that is learned. From cultural anthropology more than from any other single social science has come the emphasis on the cultural foundations of man's behaviour and thought in society.

Scarcely less than political science or economics, cultural anthropology shared in the themes of the two revolutions and their impact on the world. If the data that cultural anthropologists actually worked with were generally in the remote areas of the world, it was the effects of the two revolutions that, in a sense, kept opening up these parts of the world to more and more systematic inquiry. And, as was true of the other social sciences, the cultural anthropologists were immersed in problems of economics, polity, social class, and community, albeit among preliterate rather than "modern" peoples.

Overwhelmingly, without major exception indeed, the science of cultural anthropology was evolutionary in thrust in the 19th century. Edward B. Tylor and Sir John Lubbock in England, Lewis Henry Morgan in the United States, Adolf Bastian and Theodor Waitz in Germany, and all others in the main line of the study of primitive culture saw existing native societies in the world as prototypes of their own "primitive ancestors," fossilized remains, so to speak, of stages of development that western Europe had once gone through. Despite the vast array of data compiled on non-Western cultures, the same basic European-centred objectives are to be found among cultural anthropologists as among other social scientists in the century. Almost universally, then, the modern West was regarded as the latest point in a line of progress that was single and unilinear and on which all other peoples in the world could be fitted as illustrations, as it were, of Western man's own past.

Sociology.

Sociology came into being in precisely these terms, and during much of the century it was not easy to distinguish between a great deal of so-called sociology and social or cultural anthropology. Even if almost no sociologists in the century made empirical studies of primitive peoples, as did the anthropologists, their interest in the origin, development, and probable future of mankind was not less great than what could be found in the writings of the anthropologists. It was Auguste Comte who coined the word sociology, and he used it to refer to what he imagined would be a single, all-encompassing, science of society that would take its place at the top of the hierarchy of sciences - a hierarchy that Comte saw as including astronomy (the oldest of the sciences historically) at the bottom and with physics, chemistry, and biology rising in that order to sociology, the latest and grandest of the sciences. There was no thought in Comte's mind - nor was there in the mind of Herbert Spencer, whose general view of sociology was very much like Comte's - of there being other, competing social sciences. Sociology would be to the whole of the social world what each of the other great sciences was to its appropriate sphere of reality.

Both Comte and Spencer believed that civilization as a whole was the proper subject of sociology. Their works were concerned, for the most part, with describing the origins and development of civilization and also of each of its major institutions. Both declared sociology's main divisions to be "statics" and "dynamics," the former concerned with processes of order in society, the latter with processes of evolutionary change in society. Both men also saw all existing societies in the world as reflective of the successive stages through which Western society had advanced in time over a period of tens of thousands of years.

Not all sociologists in the 19th century conceived their discipline in this light, however. Side by side with the "grand" view represented by Comte and Spencer were those in the century who were primarily interested in the social problems that they saw around them - consequences, as they interpreted them, of the two revolutions, the industrial and democratic. Thus in France just after midcentury, Frédéric Le Play published a monumental study of the social aspects of the

working classes in Europe, Les Ouvriers européens, which compared families and communities in all parts of Europe and even other parts of the world. Alexis de Tocqueville, especially in the second volume of his Democracy in America (1835), provided an account of the customs, social structures, and institutions in America, dealing with these - and also with the social and psychological problems of Americans in that day - as aspects of the impact of the democratic and industrial revolutions upon traditional society.

At the very end of the 19th century, in both France and Germany, there appeared some of the works in sociology that were to prove most lasting in their effects upon 20th-century sociology. Ferdinand Tönnies, in his Gemeinschaft und Gesellschaft (1887; translated as Community and Society), sought to explain all major social problems in the West as the consequence of the West's historical transition from the communal, status-based, concentric society of the Middle Ages to the more individualistic, impersonal, and large-scale society of the democratic-industrial period. In general terms, allowing for individual variations of theme, these were the views of Max Weber, Georg Simmel, and Émile Durkheim (all of whom also wrote in the late 19th and early 20th century). These were the men who, starting from the problems of Western society that could be traced to the effects of the two revolutions, did the most to establish the discipline of sociology as it is found for the most part in the 20th century.

Social psychology.

Social psychology as a distinct discipline also originated in the 19th century, although its outlines were perhaps somewhat less clear than was true of the other social sciences. The close relation of the human mind to the social order, its dependence upon education and other forms of socialization, was well known in the 18th century. In the 19th century, however, an ever more systematic discipline came into being to uncover the social and cultural roots of human psychology and also the several types of "collective mind" that analysis of different cultures and societies in the world might reveal. In Germany, Moritz Lazarus and Wilhelm

Wundt sought to fuse the study of psychological phenomena with analyses of whole cultures. Folk psychology, as it was called, did not, however, last very long in scientific esteem.

Much more esteemed, and closer to 20th-century conceptions of social psychology, were the works of such men as Gabriel Tarde, Gustave Le Bon, Lucien Lévy-Bruhl, and Émile Durkheim in France and Georg Simmel in Germany (all of whom also wrote in the early 20th century). Here, in concrete, often highly empirical studies of small groups, associations, crowds, and other aggregates (rather than in the main line of psychology during the century, which tended to be sheer philosophy at one extreme and a variant of physiology at the other) are to be found the real beginnings of social psychology. Although the point of departure in each of the studies was the nature of association, they dealt, in one degree or other, with the internal processes of psychosocial interaction, the operation of attitudes and judgments, and the social basis of personality and thought - in short, with those phenomena that would, in the 20th century, be the substance of social psychology as a formal discipline.

Social statistics and social geography.

Two final manifestations of the social sciences in the 19th century are social statistics and social (or human) geography. At that time, neither achieved the notability and acceptance in colleges and universities that such fields as political science and economics did. Both, however, were as clearly visible by the latter part of the century as any of the other social sciences. And both were to exert a great deal of influence on the other social sciences by the beginning of the 20th century: social statistics on sociology and social psychology pre-eminently; social geography on political science, economics, history, and certain areas of anthropology, especially those areas dealing with the dispersion of races and the diffusion of cultural elements. In social statistics the key figure of the century was a Belgian, Adolphe Quetelet, who was the first, on any systematic basis, to call attention to the kinds of structured behaviour that could be observed and identified

only through statistical means. It was Quetelet who brought into prominence the momentous concept of "the average man" and his behaviour. The two major figures in social or human geography in the century were Friedrich Ratzel in Germany and Paul Vidal de la Blache in France. Both broke completely with the crude environmentalism of earlier centuries, which had sought to show how topography and climate actually determine human behaviour, and they substituted the more subtle and sophisticated insights into the relationships of land, sea, and climate on the one hand and, on the other, the varied types of culture and human association that are to be found on earth.

In summary, by the end of the 19th century all the major social sciences had achieved a distinctiveness, an importance widely recognized, and were, especially in the cases of economics and political science, fully accepted as disciplines in the universities. Most important, they were generally accepted as sciences in their own right rather than as minions of philosophy.

THE 20TH CENTURY

What is seen in the 20th century is not only an intensification and spread of earlier tendencies in the social sciences but also the development of many new tendencies that, in the aggregate, make the 19th century seem by comparison one of quiet unity and simplicity in the social sciences.

In the 20th century, the processes first generated by the democratic and industrial revolutions have gone on virtually unchecked in Western society, penetrating more and more spheres of once traditional morality and culture, leaving their impress on more and more nations, regions, and localities. Equally important, perhaps in the long run far more so, is the spread of these revolutionary processes to the non-Western areas of the world. The impact of industrialism, technology, secularism, and individualism upon peoples long accustomed to the ancient unities of tribe, local community, agriculture, and religion was first to be seen in the context of colonialism, an outgrowth of nationalism and capitalism in the West. The relations

of the West to non-Western parts of the world, the whole phenomenon of the "new nations," are vital aspects of the social sciences.

So too are certain other consequences, or lineal episodes, of the two revolutions. The 20th century is the century of nationalism, mass democracy, and large-scale industrialism beyond reach of any 19th-century imagination so far as magnitude is concerned. It is the century of mass warfare, of two world wars with toll in lives and property greater perhaps than the sum total of all preceding wars in history. It is the century too of totalitarianism: Communist, Fascist, and Nazi; and of techniques of terrorism that, if not novel, are to be seen on a scale and with an intensity of scientific application that could scarcely have been predicted by those who considered science and technology as unqualifiedly humane in possibility. It is a century of affluence in the West, without precedent for the masses of people, to be seen in a constantly rising standard of living and a constantly rising level of expectations.

The last is important. A great deal of the turbulence in the 20th century - political, economic, and social - is the result of desires and aspirations that have been constantly escalating and that have been passing from the white people in the West to ethnic and racial minorities among them and, then, to whole continents elsewhere. Of all manifestations of revolution, the revolution of rising expectations is perhaps the most powerful in its consequences. For, once this revolution gets under way, each fresh victory in the struggle for rights, freedom, and security tends to magnify the importance of what has not been won.

Once it was thought that, by solving the fundamental problems of production and large-scale organization, man could ameliorate other problems, those of a social, moral, and psychological nature. What in fact occurred, on the testimony of a great deal of the most notable thought and writing, was a heightening of such problems. It would appear that as man satisfies, relatively at least, the lower order needs of food and shelter, his higher order needs for purpose and meaning in life become ever more imperious. Thus such philosophers of history as Arnold Toynbee, Pitirim Sorokin, and Oswald Spengler have dealt with problems of purpose and

meaning in history with a degree of learning and intensity of spirit not seen perhaps since St. Augustine wrote his monumental The City of God in the early 5th century when signs of the disintegration of Roman civilization were becoming overwhelming in their message to so many of that day. In the 20th century, though the idea of progress has certainly not disappeared, it has been rivalled by ideas of cyclical change and of degeneration of society. It is hard to miss the currency of ideas in modern times - status, community, purpose, moral integration, on the one hand, and alienation, anomie, disintegration, breakdown on the other - that reveal only too clearly the divided nature of man's spirit, the unease of his mind.

There is to be seen too, especially during later decades of the century, a questioning of the role of reason in human affairs - a questioning that stands in stark contrast with the ascendancy of rationalism in the two or three centuries preceding. Doctrines and philosophies stressing the inadequacy of reason, the subjective character of human commitment, and the primacy of faith have rivalled - some would say conquered - doctrines and philosophies descended from the Age of Reason. Existentialism, with its emphasis on the basic loneliness of the individual, on the impossibility of finding truth through intellectual decision, and on the irredeemably personal, subjective character of man's life, has proved to be a very influential philosophy in the writings of the 20th century. Freedom, far from being the essence of hope and joy, is the source of man's dread of the universe and of his anxiety for himself. Søren Kierkegaard's 19th-century intimations of anguished isolation as the perennial lot of the individual have had rich expression in the philosophy and literature of the 20th century.

It might be thought that such intimations and presentiments as these have little to do with the social sciences. This is true in the direct sense perhaps but not true when one examines the matter in terms of contexts and ambiences. The "lost individual" has been of as much concern to the social sciences as to philosophy and literature. Ideas of alienation, anomie, identity crisis, and estrangement from norms are rife among the social sciences, particularly, of course, those most directly concerned with the nature of the social bond, such as sociology, social psychology,

and political science. In countless ways, interest in the loss of community, in the search for community, and in the individual's relation to society and morality have had expression in the work of the social sciences. Between the larger interests of a culture and the social sciences there is never a wide gulf - only different ways of defining and approaching these interests.

Marxist influences.

The influence of Marxism in the 20th century must not be missed. Currently the works of Lenin have outstripped the Bible in distribution in the world. For hundreds of millions of persons today the ideas of Marx, as communicated by Lenin, have profound moral, even religious, significance. But even in those parts of the world, the West foremost, where Communism has exerted little direct political impact, Marxism remains a potent source of ideas. Not a few of the central concepts of social stratification and the location and diffusion of power in the social sciences come straight from Marx's insights. Far more was this the case in the Communist countries - the former Soviet Union, other eastern European countries, China, and even Asian countries in which no Communist domination exists. In all these countries, Marx's name is virtually sacrosanct. There is not the same degree of differentiation of social sciences in these countries that is found in the West. As an example, sociology hardly exists as a recognized discipline in these countries, and, by the standards of the West, the other social sciences have little more than a rather rudimentary existence. Economics alone tends to be favoured, and this is, of course, largely Marxian economics - the economics of Marx's Das Kapital.

But, though Marxism has had relatively little direct impact on the social sciences as disciplines in the West, it has had enormous influence on states of mind that are closely associated with the social sciences. Especially was this true during the 1930s, the decade of the Great Depression. Today signs are not lacking of a strong revival of interest in Marx that could well, through sheer numbers of its adherents, affect the nature of the social sciences in the years ahead. Socialism remains for

many an evocative symbol and creed. Marx remains a formidable name among intellectuals and is still, without any question, the principal intellectual source of radical movements in politics. Such a position cannot help but influence the contexts of even the most abstract of the social sciences.

What Marx's ideas have suggested above all else in a positive way is the possibility of a society directed not by blind forces of competition and struggle among economic elements but instead by directed planning. This hope, this image, has proved a dominant one in the 20th century even where the influence of Marx and of Socialism has been at best small and indirect. It is this profound interest in central planning and governance that has given almost historic significance to the ideas of the English economist J.M. Keynes. What is called Keynesianism has as its intellectual base a very complex modification of the classical doctrines of economics - one set forth in Keynes's famous The General Theory of Employment, Interest and Money, published in 1935-36. Of greater influence today, however, than the strictly theoretical content of this general theory is the political impact that Keynesian ideas have had on Western democracies. For out of these ideas came the clear policy of governments dealing directly with the business cycle, of pumping money and credit into an economic system when the cycle threatens to turn downward, and of then lessening this infusion when the cycle moves upward. Above all other names in the West, that of Keynes has become identified with such policy in the democracies and with the general movement of central governments toward ever more active and constant regulation of processes once thought best left to what the classical economists thought of as natural laws. True, the root ideas of the classical economists are found in modified form even today in the works of such economists as the American Milton Friedman. But it would not be unfair to say that Keynes's name has become associated with democratic economic planning and direction in much the way that Marx's name is associated with Communist economic policies.

Freudian influences.

In the general area of personality, mind, and character, the writings of Sigmund Freud have had influence on 20th-century culture and thought scarcely less than Marx's. His basic theories of the role of the unconscious mind, of the lasting effects of infantile sexuality, and of the Oedipus complex have gone beyond the discipline of psychoanalysis and even the larger area of psychiatry to areas of several of the social sciences. Anthropologists have applied Freudian concepts to their studies of primitive cultures, seeking to assess comparatively the universality of states of the unconscious that Freud and his followers held to lie in the whole human race. Some political scientists have used Freudian ideas to illuminate the nature of authority generally, and political power specifically, seeing in totalitarianism, for example, the thrust of a craving for the security that total power can give. Sociology and social psychology have been influenced by Freudian ideas in their studies of social interaction and motivation. From Freud came the fruitful perspective that sees social behaviour and attitudes as generated not merely by the external situation but also by internal emotional needs springing from childhood - needs for recognition, authority, self-expression. Whatever may be the place directly occupied by Freud's ideas in the social sciences today, his influence upon 20th-century thought and culture generally, not excluding the social sciences, has been hardly less than Marx's.

Specialization and cross-disciplinary approaches.

A major point to make about the social sciences of the 20th century is the vast increase in the number of social scientists involved, in the number of academic and other centres of teaching and research in the social sciences, and in their degree of both comprehensiveness and specialization. The explosion of the sciences generally in the 20th century - an explosion responsible for the fact that a majority of all scientists who have ever lived in human history are now alive - has had, as one of its signal elements, the explosion of the social sciences. Not only has there been development and proliferation but there has also been a spectacular diffusion of the social sciences. Beginning in a few places in western Europe and the United

States in the 19th century, the social sciences, as bodies of ongoing research and centres of teaching, are today to be found almost everywhere in the world. In considerable part this has followed the spread of universities from the West to other parts of the world and, within universities, the very definite shift away from the hegemony once held by humanities alone to the near-hegemony held today by the sciences, physical and social.

Specialization has been as notable a tendency in the social sciences as in the biological and physical sciences. This is reflected not only in varieties of research but also in course offerings in academic departments. Whereas not very many years ago, a couple of dozen advanced courses in a social science reflected the specialization and diversity of the discipline even in major universities with graduate schools, today a hundred such courses are found to be not enough.

Side by side with this strong trend toward specialization, however, is another, countering trend: that of cross-fertilization and interdisciplinary cooperation. At the beginning of the century, down in fact until World War II, the several disciplines existed each in a kind of splendid isolation from the others. That historians and sociologists, for example, might ever work together in curricula and research projects would have been scarcely conceivable prior to about 1945. Each social science tended to follow the course that emerged in the 19th century: to be confined to a single, distinguishable, if artificial, area of social reality. Today, evidences are all around of cross-disciplinary work and of fusion within a single social science of elements drawn from other social sciences. Thus there are such vital areas of work as political sociology, economic anthropology, psychology of voting, and industrial sociology. Single concepts such as "structure," "function," "alienation," and "motivation" can be seen employed variously to useful effect in several social sciences. The techniques of one social science can be seen consciously incorporated into another or into several social sciences. If history has provided much in the way of perspective to sociology or anthropology, each of these two has provided perspective, and also whole techniques, such as statistics

and survey, to history. In short, specialization is by no means without some degree at least of countertendencies such as fusion and synthesis.

Another outstanding characteristic of each of the social sciences in the 20th century is its professionalization. Without exception, the social sciences have become bodies of not merely research and teaching but also practice, in the sense that this word has in medicine or engineering. Down until about World War II, it was a rare sociologist or political scientist or anthropologist who was not a holder of academic position. There were economists and psychologists to be found in banks, industries, government, even in private consultantship, but the numbers were relatively tiny. Overwhelmingly the social sciences had visibility alone as academic disciplines, concerned essentially with teaching and with more or less basic, individual research. All this has changed profoundly, and on a vast scale, during the past three decades.

Today there are as many economists and psychologists outside academic departments as within, if not more. The number of sociologists, political scientists, and demographers to be found in government, industry, and private practice rises constantly. Equally important is the changed conception or image of the social sciences. Today, to a degree unknown before World War II, the social sciences are conceived as policy-making disciplines, concerned with matters of national welfare in their professional capacities in just as sure a sense as any of the physical sciences. Inevitably, tensions have arisen within the social sciences as the result of processes of professionalization. Those persons who are primarily academic can all too easily feel that those who are primarily professional have different and competing identifications of themselves and their disciplines.

Nature of the research.

The emphasis upon research in the social sciences has become almost transcending within recent decades. This situation is not at all different from that which prevails in the physical sciences and the professions in this age. Prior to about 1945, the functions of teaching and research had approximately equal value in many

universities and colleges. The idea of a social (or physical) scientist appointed to an academic institution for research alone, or with research preponderant, was scarcely known. Research bureaus and institutes in the social sciences were very few and did not rival traditional academic departments and colleges as prestige-bearing entities. All of that was changed decisively beginning with the period just after World War II. From governments and foundations, large sums of money passed into the universities - usually not to the universities as such, but rather to individuals or small groups of individuals, each eminent for research. Research became the uppermost value in the social sciences (as in the physical) and hence, of course, in the universities themselves.

Probably the greatest single change in the social sciences during the past generation has been the widespread introduction of mathematical and other quantitative methods. Without question, economics is the discipline in which the most spectacular changes of this kind have taken place. So great is the dominance of mathematical techniques here - resulting in the eruption of what is called econometrics to a commanding position in the discipline - that, to the outsider, economics today almost appears to be a branch of mathematics. But in sociology, political science, social psychology, and anthropology, the impact of quantitative methods, above all, of statistics, has also been notable. No longer does statistics stand alone, a separate discipline, as it did in effect during the 19th century. This area today is inseparable from each of the social sciences, though, in the field of mathematics, statistics still remains eminently distinguishable, the focus of highly specialized research and theory.

Within the past decade or two, the use of computers and of all the complex techniques associated with computers has become a staple of social-science research and teaching. Through the data storage and data retrieval of electronic computers, working with amounts and diversity of data that would call for the combined efforts of hundreds, even thousands of technicians, the social sciences have been able to deal with both the extensive and intensive aspects of human behaviour in ways that would once have been inconceivable. The so-called

computer revolution in modern thought has been, in short, as vivid a phase of the social as the physical sciences, not to mention other areas of modern life. The problem as it is stated by mature social scientists is to use computers in ways in which they are best fitted but without falling into the fallacy that they can alone guide, direct, and supply vital perspective in the study of man.

Closely related to mathematical, computer, and other quantitative aspects of the social sciences is the vast increase in the empiricism of modern social science. Never in history has so much in the way of data been collected, examined, classified, and brought to the uses of social theory and social policy alike. What has been called the triumph of the fact is nowhere more visible than in the social sciences.

Without question, this massive empiricism has been valuable, indispensable indeed, to those seeking explanations of social structures and processes. Empiricism, however, like quantitative method, is not enough in itself. Unless related to hypothesis, theory, or conclusion, it is sterile, and most of the leading social scientists of today reflect this view in their works. Too many, however, deal with the gathering and classifying of data as though these were themselves sufficient.

It is the quest for data, for detailed, factual knowledge of human beliefs, opinions, and attitudes, as well as patterns and styles of life - familial, occupational, political, religious, and so on - that has made the use of surveys and polls another of the major tendencies in the social sciences of this century. The poll data one sees in his newspaper are hardly more than the exposed portion of an iceberg. Literally thousands of polls, questionnaires, and surveys are going on at any given moment today in the social sciences. The survey or polling method ranks with the quantitative indeed in popularity in the social sciences, both being, obviously, indispensable tools of the empiricism just mentioned.

Theoretical modes.

It is not the case, however, that interest in theory is a casualty of the 20th-century fascination with method and fact. Though there is a great deal less of that grand or

comprehensive theory that was a hallmark of 19th-century social philosophy and social science, there are still those persons occasionally to be found today who are engrossed in search for master principles, for general and unified theory that will assimilate all the lesser and more specialized types of theory. But their efforts and results are not regarded as successful by the vast majority of social scientists. Theory, at its best, today tends to be specific theory - related to one or other of the major divisions of research within each of the social sciences. The theory of the firm in economics, of deviance in sociology, of communication in political science, of attitude formation in social psychology, of divergent development in cultural anthropology are all examples of theory in every proper sense of the word. But each is, clearly, specific. If there is a single social science in which a more or less unified theory exists, with reference to the whole of the discipline, it is economics. Even here, however, unified, general theory does not have the sovereign sweep it had in the classical tradition of Ricardo and his followers before the true complexities of economic behaviour had become revealed.

Developmentalism.

Developmentalism is another overall influence upon the work of the social sciences, especially within the past three decades. As noted above, an interest in social evolution was one of the major aspects of the social sciences throughout the 19th century in western Europe. In the early 20th century, however, this interest, in its larger and more visible manifestations, seemed to terminate. There was a widespread reaction against the idea of unilinear sequences of stages, deemed by the 19th-century social evolutionists to be universal for all mankind in all places. Criticism of social evolution in this broad sense was a marked element of all the social sciences, pre-eminently in anthropology but in the others as well. There were numerous demonstrations of the inadequacy of unilinear descriptions of change when it came to accounting for what actually happened, so far as records and other evidences suggested, in the different areas and cultures of the world.

Beginning in the late 1940s and the 1950s, however, there was a resurgence of developmental ideas in all the social sciences - particularly with respect to studies of the new nations and cultures that were coming into existence in considerable numbers. Studies of economic growth and of political and social development have become more and more numerous. Although it would be erroneous to see these developmental studies as simple repetitions of those of the 19th-century social evolutionists, there are, nevertheless, common elements of thought, including the idea of stages of growth and of change conceived as continuous and cumulative and even as moving toward some more or less common end. At their best, these studies of growth and development in the new nations, by their counterposing of traditional and modern ways, tell a good deal about specific mechanisms of change, the result of the impact of the West upon outlying parts of the world. But as more and more social scientists have recently become aware, efforts to place these concrete mechanisms of change into larger, more systematic models of development all too commonly succumb to the same faults of unilinearity and specious universalism that early-20th-century critics found in 19th-century social evolution.

Social-systems approach.

Still another major tendency in all of the social sciences since World War II has been the interest in "social systems." The behaviour of individuals and groups is seen as falling into multiple interdependencies, and these interdependencies are considered sufficiently unified to warrant use of the word "system." Although there are clear uses of biological models and concepts in social-systems work, it may be fair to say that the greatest single impetus to development of this area was widening interest after World War II in cybernetics - the study of human control functions and of the electrical and mechanical systems that could be devised to replace or reinforce them. Concepts drawn from mechanical and electrical engineering have been rather widespread in the study of social systems.

In social-systems studies, the actions and reactions of individuals, or even of groups as large as nations, are seen as falling within certain definable, more or less universal patterns of equilibrium and disequilibrium. The interdependence of roles, norms, and functions is regarded as fundamental in all types of group behaviour, large and small. Each social system, as encountered in social-science studies, is a kind of "ideal type," not identical to any specific "real" condition but sufficiently universal in terms of its central elements to permit useful generalization.

Structuralism and functionalism.

Structuralism in the social sciences is closely related to the theory of the social system. Although there is nothing new about the root concepts of structuralism - they may be seen in one form or other throughout Western thought - there is no question but that in the present century this view of behaviour has become a dominant one in many fields. At bottom it is a reaction against all tendencies to deal with human thought and behaviour atomistically - that is, in terms of simple, discrete units of either thought, perception, or overt behaviour. In psychology, structuralism in its oldest sense simply declares that perception occurs, with learning following, in terms of experiences or sensations in various combinations, in discernible patterns or gestalten. In sociology, political science, and anthropology, the idea of structure similarly refers to the repetitive patternings that are found in the study of social, economic, political, and cultural existence. The structuralist contends that no element can be examined or explained outside its context or the pattern or structure of which it is a part. Indeed, it is the patterns, not the elements, that are the only valid objects of study.

What is called functionalism in the social sciences today is closely related to structuralism, with the term structural-functional a common one, especially in sociology and anthropology. Function refers to the way in which behaviour takes on significance, not as a discrete act but as the dynamic aspect of some structure. Biological analogies are common in theories of structure and function in the social sciences.

Very common is the image of the biological organ, with its close interdependence to other organs (as the heart to the lung) and the interdependence of activities (as circulation to respiration).

Interactionism.

Interaction is still another concept that has had wide currency in the social sciences of the 20th century. Social interaction - or, as it is sometimes called, symbolic interaction - refers to the fact that the relationships among two or more groups or human beings are never one-sided, purely physical, or direct. Always there is reciprocal influence, a mutual sense of "otherness." And always the presence of the "other" has crucial effect in one's definition of not merely what is external but what is internal. One acquires one's individual sense of identity from interactions with others beginning in infancy. It is the initial sense of the other person - mother, for example - that in time gives the child its sense of self, a sense that requires continuous development through later interactions with others. From the point of view of interactionist theory, all one's perceptions of and reactions to the external world are mediated or influenced by prior ideas, valuations, and assessments. Always one is engaged in socialization or the modification of one's mind, role, and behaviour through contact with others.

FUTURE OF THE SOCIAL SCIENCES

What has been covered in the preceding paragraphs may be the most that can be said within restricted compass about the social sciences of the 20th century without turning to the individual social sciences themselves and related disciplines. The concern here has been with only those major contextual influences, tendencies of overall character, and dominant ideas or theories that the social sciences taken as a whole manifest in one degree or other.

There is one final aspect of the subject that must be considered briefly, for how it is resolved will have much effect upon the future of the social sciences in the West. This is the relation of the social sciences to organized society, to government and

industry, and other institutional centres of authority. At the present time, there is a significant and undoubtedly growing feeling among social scientists, especially younger ones, that the relationship has become altogether too close. The social sciences, it is said, must maintain their distance, their freedom, from bureaucratized government and industry. Otherwise they will lose their inherent powers of honest and dispassionate criticism of the ineffective or evil in society. Although there may be a certain amount of feeling ranging from the naïve to the politically revolutionary in such sentiments, they cannot be taken lightly, as is apparent from the serious consideration that is being given on a steadily rising scale to the whole problem of the relationship between social science and social policy.

Since the inception of the social sciences - since, indeed, the time when the universities in the West came into being for the express purpose of training professional men in law, theology, and medicine – man has properly sought, through knowledge, to influence social policy, taking this latter term in the widest sense to include not merely the policies of national government but of local government, business, professions, and so on. What else, it may be asked, are the social sciences all about if it is not to use knowledge to improve social life; and how else but through influencing of the major institutions can such improvement take place?

So much is true, comes the answering response. But in the process of seeking to influence the great agencies of modern power and function - of what is loosely called the Establishment - the social sciences may themselves become influenced adversely by the values of power and affluence to be found in these great agencies. They themselves may become identified with the status quo. What the social sciences should give, say the partisans of this view, is a continuation of the revolutionary or at least profoundly reformist tradition that was begun in the 18th century by the philosophers of reason who, detesting the official establishment of their day, sought on their own to transform it. What is today called objectivity or

methodological rigour turns out to be, say these same partisans, acceptance of the basic values of reigning government and industry.

It is this essential conflict regarding the purposes of the social sciences, the relation of the social sciences to government and society, and the role of the individual social scientist in the society of the 20th century that bids fair at this moment to be the major conflict of the years ahead. How it is resolved may very well determine the fate of the social sciences, now less than two centuries old.

Sociology

Sociology is a branch of the science of human behaviour that seeks to discover the causes and effects that arise in social relations among persons and in the intercommunication and interaction among persons and groups. It includes the study of the customs, structures, and institutions that emerge from interaction, of the forces that hold together and weaken them, and of the effects that participation in groups and organizations have on the behaviour and character of persons. Sociology is also concerned with the basic nature of human society, locally and universally, and with the various processes that preserve continuity and produce change.

It is social life that is distinctive in the regulation of behaviour in human beings; the human animal does not have such instincts as serve to guide the behaviour of lower animals, and he is therefore more dependent on social organization than is any other species. Institutionalized social forms therefore are assumed to play the major part in influencing human actions, and it is the task of sociology to discover how these forms operate on the person, as well as how they are established, develop, elaborate, interact with one another, and decay and disappear. Among the most important of such structures is the family, the subject of an important field of sociology. The peer group, the community, the economic and political orders, various voluntary associations, and special organizations such as the church and the military are of particular importance in this inquiry.

Though sociology can be considered as a part of the Western tradition of rational inquiry inaugurated by the ancient Greeks, it is specifically the offspring of 18th-

and 19th-century philosophy and has been viewed as a reaction against the frequently nonscientific approaches of classical philosophy and folklore to social phenomena. It was for a time presented as a part of moral philosophy, which covered the subject matter that eventually also became the concern of the various social sciences that are now separate from moral philosophy. Some aspects of other fields remain of interest to the sociologist. Although psychology has traditionally centred its interest on the individual and his internal mental mechanisms, and although sociology has given its major attention to collective aspects of human behaviour, the two disciplines share the subfield of social psychology. The relation of sociology to social anthropology is even closer, and until about the first quarter of the 20th century the two subjects were usually combined in one department, differentiated mainly by the emphasis of the anthropologists on the sociology of preliterate peoples. Recently even this distinction has been fading, as social anthropologists have increasingly added studies of various aspects of modern society to their field of interest. Political science and economics had much of their early development in the practical interests of nations and for a time evolved separately from basic sociology; but recently in both fields an awareness of the potential utility of some infusion of sociological concepts and methods has brought relations closer. A somewhat similar situation has also been developing in respect to law, education, and religion and to a lesser extent in such contrasting fields as engineering and architecture.

Nineteenth-century sociology, influenced by the successes of biology and evolutionary theory, took an interest in resemblances between men and lower animals - in their having, for example, similar instincts - and also in the parallels between biological and social evolution. These interests have declined, but sociology continues to share with the other sciences some interest in ecology, behavioral genetics, and questions of fertility and mortality as they relate to population studies. There is also a conviction among sociologists that contact between physiology and sociology is necessary to avoid errors of ignorance in both fields.

HISTORICAL DEVELOPMENT OF SOCIOLOGY

Early major schools of thought.

The founders of sociology spent decades almost exclusively in the process of finding a direction for their new discipline. In the course of this groping effort they tried several highly divergent pathways, some suggested by methods and contents of other sciences, others invented outright by the imagination of the scholar.

Social Darwinism and evolutionism.

Darwinian evolutionary theory doubtlessly suggested a way in which a science of human behaviour could become academically respectable, and a line of creative thinkers, including Herbert Spencer, Benjamin Kidd, Lewis H. Morgan, E.B. Tylor, L.T. Hobhouse, and others, developed analogies between human society and the biological organism and introduced into sociological theory such biological concepts as variation, natural selection, and inheritance - evolutionary factors resulting in the progress of societies through stages of savagery and barbarism to civilization, by virtue of the survival of the fittest.

Some writers also perceived in the growth stages of each individual a recapitulation of these stages of society. Strange customs were thus accounted for on the assumption that they were throwbacks to an earlier useful practice; an example offered was the make-believe struggle sometimes enacted at marriage ceremonies between the bridegroom and the relatives of the bride, reflecting an earlier bride-capture custom.

Social Darwinism waned in the 20th century, but in its popular period it was used to justify unrestricted competition and a laissez-faire doctrine in order that the "fittest" would survive and that civilization would continue to advance.

Determinism: economic, environmental, biological.

Except in the philosophy of Karl Marx (whose writings ranged over all the social science fields rather than specifically in sociology), the doctrine of economic

determinism never gained a strong foothold in sociology. This was not a consequence of scholarly ignorance; sociologists of all periods have read Marx and have usually read such writers as the historian Charles A. Beard, who emphasized economic self-interest, and Werner Sombart, the German sociologist who had been a convinced Marxist in his early career. But there have been only some adapted reflections of these economic views in the writings of such sociologists as Franklin H. Giddings or Frank H. Hankins who viewed some political and religious doctrines as rationalizations of economic and social interests.

The human geographers - Ellsworth Huntington, Ellen Semple, Friedrich Ratzel, Paul Vidal de La Blache, Jean Brunhes, and others - were also read critically by sociologists but did not make a lasting major contribution to the mainstream of sociological thought, even though there are some who believe that the social morphology of Émile Durkheim, Maurice Halbwachs, and others - that is, their theories about the roles of individuals interacting in a social system - grew in part from this interest.

Aside from the interest in evolution, organismic analogies, and the instinct concept, sociologists have not found biological determination of value to them and have spent more energy in refuting it than in making use of it.

Early functionalism.

Following the achievement of a consensus that there should be a place for a science of sociology, there emerged an international effort to define the distinctive character of the subject and especially to clarify its differences from psychology and biology, fields that had also begun to generalize about human behaviour. A Frenchman, Émile Durkheim (1858-1917), was prominent among scholars who considered this question; he argued that there can arise from various kinds of interaction among individuals certain new properties (sui generis) not found in separate individuals. These "social facts" as he called them - collective sentiments, customs, institutions, nations - call for study and explanation on a distinctly sociological level rather than on the level of individual psychology. Furthermore,

the interrelations of the parts of a society were perceived as cohering into a unity, an integrated system with a life character of its own, exterior to the individual, and exercising constraint over his behaviour. This direction of causation, from group to individual (rather than the reverse as conceived by most biologists of the time) gave encouragement to the scholar of the new science. Some writers have designated such a view "functionalism," although the term has in recent years acquired some broader variations of meaning.

Durkheim also pointed out that groups could be held together on two contrasting bases: the sentimental attraction of similarities (mechanical solidarity), such as occurs in friendship groups and among relatives and neighbours, and the organization of complementary differences (organic solidarity), such as occurs in industrial, military, governmental, and other organizations that exist because they have tasks to perform. Other theorists of Durkheim's period, notably Henry Maine and Ferdinand Tönnies, made similar distinctions in different terms - status and contract (Maine) and Gemeinschaft and Gesellschaft (Tönnies) - and conceived of the major trend of civilization as an expansion of the latter and a relative decline of the former.

Some later anthropologists, especially Bronislaw Malinowski and A.R. Radcliffe-Brown, developed a doctrine also called functionalism, based on the recognition of the interrelatedness of the parts of a society, in bonds so thoroughly interpenetrating that a change in any single element would tend to produce a general disturbance in the whole. This concept gained a following for a time among many social anthropologists, leading some to advocate a policy of complete noninterference with even the most objectionable practices in a preliterate society (such as headhunting) for fear that control might produce far-reaching disorganization.

William G. Sumner, in his Folkways, defined an institution as a "concept and a structure," meaning a purpose or function that is carried out by some systematic organization of persons. Much of the sociology of Max Weber consists of the analysis of societies in such terms. Georg Simmel, sometimes called the founder of

the "formal school" of sociology, viewed society as a process ("something functional") that is real and not merely an abstraction, and he built on this idea a statement of sociology consisting of a systematic analysis of social forms.

Modern major directions of interest.

The early schools of thought - each presenting a systematic formulation of sociology that implied possession of exclusive truth and that involved a conviction of the need to destroy rival systems - in time gave way to distinguishable directions of interest and emphasis that did not have to be considered inharmonious. These new directions have no dominant leaders and no clearly defined borderlines.

Functionalism and structuralism.

Following the main contributions in the earlier theoretical formulations of Charles H. Cooley, such later authors as Pitirim A. Sorokin, Talcott Parsons, Robert Merton, Everett C. Hughes, and others have elaborated on the nature of organizations and their relation to the behaviour of persons and have attempted to build workable conceptualizations of very large social systems, nations, and societies. Sorokin designated his viewpoint as "integralist" and wrote at length about the civilization-cultures that in their balance of values and conditions could be viewed as entities that had distinguishable life cycles, with "ideational," "idealistic," and "sensate" stages marking their growth and decline, thus following a philosophy-of-history tradition shared by Edward Gibbon, Oswald Spengler, and Arnold Toynbee.

Talcott Parsons has given attention to social systems in a more analytical way, inquiring into the conditions that each system must meet in order to survive (the "functional prerequisites"), the character of the standardized and stable interpersonal arrangements (structures) needed to make each system work, the relations to environmental conditions, problems of boundaries, the recruitment and control of members, and the like. Along with Robert Merton and others, he also worked on the classifications of such structures and on distinctions of function.

The subject matter and methods involved in such structural-functional analysis have indeed become so broad that some authors (such as Marion Levy) have held that it becomes synonymous with scientific analysis in general, or at least with scientific study of the nature of organization.

On a smaller scale, Kurt Lewin and his co-workers pursued somewhat parallel questions, investigating the nature of small groups, families, professional and military units, looking for arrangements and relationships of the parts of each person's "psychological life space" and of the interrelations of these to a "social space" or society's total range of action. The choice of such relatively small units for research made fruitful experimentation possible, and from Lewin's leadership grew the influential research movement that became known as group dynamics. Some writers have also applied the descriptive term microfunctionalist to this tradition.

Symbolic interactionism.

Sociologists did not for long find the 19th-century instinctivist psychology congenial, and most of them also failed to appreciate the doctrines of classical or Watsonian behaviourism, which sought to be totally objective and experimental. One influential movement in social psychology, however, did take early root and eventually became the largest and most influential field in modern American sociology. In recent years it has become known as "symbolic interactionism," but it was under development for decades before it acquired a name.

Out of early ideas expressed by J. Mark Baldwin and William James, a group of three scholars, John Dewey, George H. Mead, and Charles H. Cooley, built the foundations of a psychology that was to become most useful to sociology. In brief, their contribution was to advance the theory that mind and self are not part of the innate equipment of the human organism but arise in experience and are constructed in a social process - that is, in a process of interaction among persons in intimate, personal communication with one another. The self, or self-concept, as developed by Mead and others, is thus essentially an internalization of aspects of

an interpersonal or social process. It exists in imagery and symbolization and is internalized and organized for each person out of his perception of how other persons conceive him. This self-concept, however inexact, fluctuating, and uncertain, nevertheless functions as a guide in social behaviour - that is, persons tend to act in order to preserve the existing or desired image of their self.

William I. Thomas, a sociologist and colleague of the philosopher Mead at the University of Chicago in the early years of the 20th century, regularly taught a course in social psychology based on Mead's conceptions. Thomas was succeeded in 1919 by Ellsworth Faris, himself a psychologist but later a member of the department of sociology, and through his work the tradition was further developed and brought into closer relation to the sociological tradition of Robert E. Park and Ernest W. Burgess, also at Chicago. In this tradition an interest in an appropriate methodology accompanied the growth of substantive knowledge; Thomas particularly emphasized the value of extensive use of personal documents, life histories, and autobiographies. In recent years interest in research on the self and self-conscious behaviour has spread widely, and is now participated in by psychologists, philosophers, and essayists, as well as by a movement within sociology called "ethnomethodology," which investigates areas of symbolic interaction by informal observation, reflection, and skilled interpretation, methods sometimes called Verstehen (understanding).

Modern determinism.

Economic determinism reflects the interest that a few early sociologists took in views of Karl Marx, such as the idea that differentiation into social classes and conflict between these classes derive from economic factors and the belief that the political system is in large part a product of such social stratification. A residue of this kind of determinism is found among the self-proclaimed "Marxian sociologists."

Perhaps the most widely read of these was C. Wright Mills, whose concept of a "power elite" has been extensively and critically examined, with varying resulting

judgments on its utility. As Mills saw it, this elite constitutes an integrated ruling group of a capitalistic economic and military system, sometimes called the military-industrial complex, exercising arbitrary power in its own interests. This particular determinism is not supported by most existing objective research, which generally finds a far more pluralistic distribution of political power.

A contrasting view of class conflicts was advocated by Karl Mannheim, who saw the cleavages as ideologically produced, as divergences in modes of thought rather than as rational perception of economic interests. Since Mannheim hoped that such conflicts could be resolved, his doctrine should not be considered fully deterministic, but it did stimulate an effort to interpret the relations between ideas and actions that came to be known as the "sociology of knowledge."

Mathematical modelism.

A variety of efforts has been made to describe and investigate behaviour mathematically, through measurement and counting and the use of mathematical models. This approach in part characterized the early "sociometry" of J.L. Moreno (although its meaning has greatly drifted and broadened in recent years), the "field theory" of Kurt Lewin, and the investigations by George K. Zipf, John Q. Stewart, and others into the relations of rank and size of political units, the frequency of word use in language, and other simple arithmetic relations. Some of the concepts of game theory, first introduced into economics by its inventors, John von Neumann and Oskar Morgenstern, have also penetrated into sociology. Also the rapidly expanding use of computers has in recent times encouraged the development of various kinds of simulation of behaviour. Some investigations of complex interaction patterns have been carried out by devising games with rules to fit the problem and persons to execute the roles. When specified rules become highly detailed and complex, the outcome may be sought through the use of a computer; thus the game is converted into a simulation. Sociologists have participated, along with other social scientists, in the creation of such simulations

of various political and military processes. Extension of these techniques into a variety of interaction processes is to be expected.

METHODOLOGICAL CONSIDERATIONS IN CONTEMPORARY SOCIOLOGY

Much of 19th-century sociology was devoid of systematic method, but late in the period the proliferation of schools of thought, based on speculative sociologies, made evident the need for ways of obtaining verifiable knowledge. Early attempts were crude and unfruitful; such broad surveyors as Charles Booth, who produced a monumental series on London, relied mainly on the gathering of masses of facts.

Frédéric Le Play in France made extensive studies of family budgets. Herbert Spencer and others assembled vast stores of observations made by other persons, using these to illustrate and support generalizations already formulated.

Early exploitation of statistical materials, such as officially recorded rates of births, deaths, crimes, and suicides, provided only a moderate advance in knowledge, because this approach was too capable of supporting preconceived ideas. Among the most successful of this type of study was research on suicide by Émile Durkheim, whose successors in France and elsewhere developed the methodology a considerable way toward scientific adequacy.

After the turn of the century, interest in, and the determination to achieve, a sociological methodology grew steadily. The Methodological Note, constituting the greater part of a volume in W.I. Thomas and Florian Znaniecki's Polish Peasant in Europe and America (5 vol., 1918-20), has been recognized as an important advance, not so much in methodology as in committing sociologists to the task of achieving it.

Significant advances toward scientific effectiveness occurred at the University of Chicago in the 1920s. Under the stimulation of Robert E. Park, Burgess, and their colleagues a series of studies of the metropolis was conducted. The spirit was inductive, and hypotheses were discovered in rather than imposed on gathered information. Large numbers of students took part in the effort and contributed to

both methods and findings. A conspicuous part of the effort consisted of mapping locations of various phenomena: land uses, residences of population categories (racial, ethnic, and occupational), residences of persons who commit various types of crimes or suicide, families becoming divorced or broken through desertion, and so forth. But along with such information on spatial distributions, data were sought by other means, including participant observation in groups and communities, gathering of life histories and case studies, assembly of relevant historical information, study of the life cycles of social movements and sects, and the like. Attention was explicitly given to the improvement of methodology in all of these efforts, to an extent approximately equal to the attention given to substantive findings. Here for the first time was developed a large-scale cooperative effort in which theory, methodology, and findings evolved together in an inductive process. The influence of this development at Chicago spread rapidly about the United States and in time influenced sociology almost everywhere it was studied in the world.

Statistics.

Statistical methods were introduced into sociology from other sciences, and virtually from the start, sociologists have found statistical measures of relationship of great value. Karl Pearson's "coefficient of correlation," for example, has been a popular as well as important statistical concept for the measurement of cause-and-effect relationships among continuous variables. This method reveals the degree of causal connection between two variables, though not necessarily the nature of the connection. In sociology there are types of data that are relevant to causal inquiry but do not have the characteristics that qualify for the Pearsonian coefficient. Thus, much development work has been done to provide other measures of association involving, for example, rankings of groups or individuals or qualitative comparisons (such as whether males and females differ systematically in specified qualities).

Factor analysis, also based on an elaboration of Pearsonian correlation, performs another valuable service to sociology. If there are a large number of variables causally intertwined in a complex way, it is possible that these variables can be reduced to a small number of factors. Fifty different tests of mental ability, for instance, may be in fact 50 different mixtures of only seven or eight dimensions of mental ability. Factor analysis involves reducing such variables to a more limited number of common factors and determining the relative importance of each factor in the original variables. The process has its imperfections and the computations are laborious, but the availability of computers has overcome the latter disadvantage, and in recent years the technique has increased in use.

These statistical methods and many others are applicable to all branches of sociology and are increasingly fruitful in transforming sociology into science. In general, the growth of statistical methods has been so rapid that the invention of new techniques has outstripped the ability of scholars to find data worthy of the devices. Thus the rate of progress in the near future may depend to a large extent on improvement in satisfactory data gathering and measurement. Methodologies of data gathering are in fact of major interest in sociology. Techniques of observation - of persons, groups, organizations, communities - have been extensively developed. Important for the same purpose are the various means of quantifying these observations, including scales of various kinds, sociometric techniques that make interrelations subject to statistical analysis, content analysis of written materials, and classification of cross-cultural information.

Experiments.

Experimental methods, once believed to be inapplicable to sociological research, were extensively applied by psychologists, first on individuals and later on groups. By the 1930s some psychologists – notably Kurt Lewin and his colleagues and also Muzafer Sherif - found means of conducting experiments on social interaction. Sociologists soon followed their example and in time a number of laboratories for such research were established; Robert F. Bales, at Harvard, has made systematic

observations on interaction in small, artificial groups and has produced clear and useful results, confirmed in other laboratories. Experiments are also conducted in classrooms, in summer camps, in formal organizations, and elsewhere. In general the success of experimentation has been greatest in simple situations in which the number of variables is limited. Complex experiments, however, are possible in some circumstances, and the design of complex formal experiments is becoming a developed art in a variety of fields, including sociology.

Data collection.

Within the main categories of research methods there are many special problems for which techniques have been devised. Data collection, for example, is effected in many different ways, from unstructured observation, essentially methodless, to sophisticated measurement through special instruments. Some of the basic problems of data collection concern such matters as the most efficient use of terminology, the definitions of units to be measured, and the classifications to be used. In general it is necessary to consider the nature of a specific problem in order to choose the most appropriate unit. For example, in a study of the relation of the size of a city to the cost of operating its local government, the proper unit might well be the population residing within its political boundaries. If the research question, however, is the relation of city size to any of a number of forms of social disorganization, it may be more fruitful to recognize that sociologically the significant unit would include much or all of the settled areas outside the city limits.

In the fields of social differentiation and occupational mobility the matter of definition of specific occupations is critical. If persons are asked in a questionnaire to state their occupation, the usual response is to give only one occupation, and this one is sometimes vaguely defined and made obscure by the tendency to give a euphemistic answer. Persons change occupations; some have more than one; some might claim an occupation that they merely aspire to. The art of obtaining useful answers to such important questions involves carefully designed questions adapted

to the specific purposes of the study. General classifications, intended for a variety of studies, have limited utility.

In the process of gathering research data for sociology there are occasional obstacles to direct observation. In such cases indirect indicators may provide crude but useful substitutes. For example, alcoholic consumption in a small village in which the beverage is supposed to be prohibited may be estimated by a count of empty bottles in trash receptacles, or perhaps in the town dump. Library book circulation has been used to estimate the use of television in a community in which withdrawals of books of fiction declined, while nonfiction withdrawals remained as before.

Questionnaires are convenient for obtaining information from large numbers of respondents but involve many methodological problems. Wording of questions must of course be intelligible to uneducated and uninterested persons, must have standard meanings to persons of varying backgrounds, must avoid topics that arouse resistance and refusal to complete the questionnaire, and must avoid being too complex or difficult so that returns are insufficient or constitute a biased sample. Since it is known that slight alterations in the wording of questionnaire items may produce considerable variations in the pattern of responses, the precise wording becomes a matter of some art as well as science. A similar effect occurs in the order of items, since some may suggest or influence responses on later ones.

Similar issues are involved in data gathering through interviewing. It is necessary to control such variables as the appearance, manner, and approach of the interviewer, the specific manner in which questions are asked, ways of avoiding interviewer influence on the responses, and the tendency of some respondents to refuse to answer questions or to discontinue the interview. To meet the problems of resistance on sensitive subjects and inarticulateness about some feelings, various indirect or projective devices may be employed so that a respondent in answering one question provides information he may not realize he is giving about other questions.

Questionnaires and interviews may be so arranged that the patterns of responses form a scale, converting qualitative variations into measures available for statistical treatment. An early scaling method, devised in the late 1920s by a psychologist, L.L. Thurstone, is still widely used in sociology. It is formed in the following way: a list of questionnaire items is presented to a numberof judges who independently relist the items in the order in which they consider them important or of interest. From their decisions are selected items on which there is satisfactory agreement of scale value.

Scaling may also be provided by statements to which a respondent is asked whether he "strongly approves," "approves," is "undecided," "disapproves," or "strongly disapproves." Or the quantitative differences may be introduced through a logical sequence of preference answers - for example, whether the respondent would admit a particular category of person (a) to close kinship by marriage, (b) to his club as a personal chum, (c) to employment in his occupation, (d) to citizenship in his country. Here it is assumed that the later answers imply more desired social distance.

A method or class of methods called sociometry has been under development since its introduction in the middle 1930s by J.L. Moreno. The essence of the method is the collection and tabulation of information about various types of interaction among members of groups of small or moderate size. The interaction may be either actual behaviour or merely anticipated or desired behaviour, and it may consist of preferences for various kinds of association with other persons, such as having them as friends, sitting with them, working with them, and the like. The information may be collected by observation of real behaviour or by interviews or questionnaires with specific items regarding personal choices. After the information is gathered, it is sometimes put in the form of a sociogram, consisting of names of persons enclosed in circles or squares distributed over an area and connected with lines and arrows that indicate both detail of choices and general patterns of relationships. A person receiving many choices is readily seen as the target end of many lines and is sometimes referred to as a "star." A person

completely unchosen has no lines pointing toward his name and is called an isolate. Further investigation of persons typed in this fashion may be made by statistical methods, case studies, or otherwise. Overall, it can be said that various improvements and elaborations of the basic sociometric approach have been made, and the method is now less distinctively separate from other social psychology research than it was originally.

Ecological patterning.

Ecological methods in sociology were first developed in connection with research on the characteristics of the metropolis, especially in regard to features of a nonsocial character, such as the patterns resulting from the distribution and movements of populations and institutions in the general process of struggling for advantage. A conspicuous part of most early urban studies consisted of mapping such distributions. The patterns of land values, of locations of various types of businesses and industries, of ethnic categories of the population, and of types of behaviour (delinquency and crime, vice, family disorganization, mental disorders, etc.) were all shown to be interrelated in a general urban ecology. This fact was then shown to be related to many aspects of behaviour of city people, and valuable contributions were made to such general sociological topics as social differentiation, migration and vertical mobility, and social disorganization.

In recent years sociological ecology has broadened in meaning and in the elaboration of methods. One modern approach, known as ecosystem theory, consists of tracing general patterns of flow of materials, energy, and information into a system and their transformation during the flow through the system, among other things.

Problems of bias.

Since most sociological knowledge is based on the study of samples from some larger universe of items, the possibilities of major errors from sampling bias constitute a methodological issue. Where biases cannot be controlled, the direction

and extent may sometimes be estimated, but elimination of biases through use of quotas - or, when possible, random methods - yields the best results. This can be done, for example, by first randomly selecting a number of definable regions and metropolitan areas, then selecting randomly from each such area certain urban blocks and rural segments, then further selecting from these segments certain dwelling units, and finally selecting from the dwelling units the specific persons to constitute the sample.

In every stage of the process of discovery in sociology there are possibilities of error, and recognition of these is a part of the progress of sociological methodology. There is continuous creation of technical devices to reduce such errors and to estimate the amount of error that has not been eliminated.

National methodological preferences.

All the methods described above are widely used, but their relative popularity in various nations is somewhat related to both the nature of the financial support of research and the field of national interest. Where agricultural problems are of major interest, rural sociology and community studies that can be conducted inexpensively by one or a few investigators are popular. In France, Italy, and several other European nations, industrial sociology is understandably important, much of it based on case studies of industries and the experiences of workers. Sociology in Great Britain, the Scandinavian countries, and Japan covers most of the fields mentioned above.

The broad methodological concepts have varied somewhat according to the country and according to the subfield of sociology. Early in the century there was presumed to be a general difference between the sociologies of European countries and the sociologies of the United States - the former appearing to prefer broad sociological theory based on philosophical methods and the latter showing more inclination toward induction and empiricism. Such differences have declined steadily in recent times, and what differences remain may be in part a result of the differential financing of expensive research.

In the former U.S.S.R. and in nations that were under its influence there was much emphasis on the concepts and methods of Marxist sociology, which had only a small following elsewhere. A more important methodological issue divides basic scientific sociology from applied sociology; scholars interested in applied sociology tend to deprecate the methods and findings of the scientific sociologists as being either irrelevant or supportive of an objectionable status quo. Issues of ethics have also in recent years been raised, particularly in regard to observations and experiments in which the privacy of subjects may be felt to be invaded.

STATUS OF CONTEMPORARY SOCIOLOGY

Professional status.

The Greek philosophers and the line of European philosophers in the succeeding centuries throughout Western civilization discussed much of the subject matter of sociology without thinking of it as a distinct subject. In the early 19th century all the subject matter of the social sciences was discussed under the heading of moral philosophy. Even after Auguste Comte introduced the word sociologie in 1838, the matter was combined with other subjects for some sixty years. Not until the universities undertook a commitment to the subject could a person make a living as a full-time sociologist. This commitment had first to be made by scholars of other fields, of which history was a principal early sponsor.

As early as 1876, at the new Johns Hopkins University, some of the content of sociology was taught in the department of history and politics. In 1889 at the University of Kansas, the word appeared in the title of the department of history and sociology. In 1890 at Colby College, a historian, Albion Small, taught a course called sociology, as did Franklin H. Giddings in the same year at Bryn Mawr College. But the first real commitment to the creation of a field of sociology took place in 1892 at the new University of Chicago, where newly arrived Albion Small asked for and received permission to create a department called sociology - the first such in the world. In the following year or two, departments in the subject were founded at Columbia, Kansas, and Michigan and very soon afterward at Yale,

Brown, and many other universities. By the late 1890s nearly all of the educational institutions in the United States either had departments of sociology or offered courses in the subject.

In 1895 the American Journal of Sociology began publication at the University of Chicago, in time to be followed by a large number of journals in many other countries. Ten years later the American Sociological Society was organized, also to be followed in time by a large number of national, regional, international, and special sociological organizations. These quickly institutionalized the subject and have continuously served to guide its directions and to establish, very roughly, its boundaries. Eventually in 1949 the International Sociological Association was established under the sponsorship of UNESCO, and Louis Wirth (1897-1952) of the University of Chicago was elected its first president.

The rapid growth in numbers of full-time sociologists, along with growth of publications, allowed the content of the discipline to expand rapidly. By 1970 there were more than a dozen important sociological journals and an indefinite number of minor journals in the U.S., as well as a considerable number in other nations. Research grew throughout the 20th century at an accelerated pace, especially since the 1920s, partly because of strong financial support from foundations, government, commercial sources, and private gifts. Along with this came a flourishing of research institutes, some affiliated with university departments and some independent. A small but increasing number of sociologists gain their livelihood through full-time research independent of universities.

Similar developments have occurred in various other parts of the world, with variations resulting from special conditions in each case. In France, where Auguste Comte and later Émile Durkheim gave early impetus to sociology, there was early development in many fields of the subject. The two world wars slowed the development, but after 1945 a strong revival of interest in sociology took place, during which the French government established a number of institutes in the social sciences at the level of institutes in the natural sciences, including several in Paris for sociological research - notably the Centre d'Études Sociologiques, the

Institut National d'Études Démographiques, and the Maison des Sciences de l'Homme. These institutes receive government funds and employ many full-time sociologists, some of them among the prominent scholars in the nation. French universities have been somewhat more conservative; the Sorbonne, for example, had in 1970 only one chair officially assigned to sociology. The new University of Nanterre, however, established a department with four professorships. A rich amount of research publication has been produced in France since World War II, particularly in general sociology, theory, methodology, social psychology, industrial sociology, and the sociology of work.

German sociology had a strong base in the late 19th century and afterward, and the writings of Ferdinand Tönnies, Max Weber, Georg Simmel, and others were influential in all parts of the world. By the early 1930s, however, official Nazi hostility had impeded its development and by the time of World War II had destroyed it as an academic subject in Germany. Immediately after the war a new generation of scholars, aided by visiting sociologists, imported the new empirical research methods and began the development of a style of German sociology much different from the earlier theoretical and philosophical traditions. At the University of Frankfurt, Max Horkheimer's Institut für Sozialforschung (social research), established by private financing before the war, was revived and has stimulated much research production. West German universities remained conservative for a time, but two newly created universities - the Free University of Berlin and the University of Constance - made sociology one of their major subjects. By 1970 most West German universities had at least one chair in sociology. National needs received special emphasis, including administrative research of use to planning, studies of unemployment, youth problems, and delinquency. A significant amount of research also is published in such fields as rural sociology, political sociology, and the family.

In Great Britain, despite the early prominence of Herbert Spencer and L.T. Hobhouse, sociology was little regarded by leading universities until the mid-20th century. Before World War II Britain excelled in anthropology, especially in the

study of nonwhite societies of the empire. Sociology concentrated on studies of the poor, and much of it was undertaken by persons whose affiliation was similar to that of social workers in the United States. The major prewar sociology department, at the London School of Economics, had the objective more of social reform than scientific research. In the postwar period, however, a considerable revival of sociology took place; Oxford and Cambridge recognized the subject by creating positions for sociologists, and various new universities established chairs and departments.

Significant work in Britain has been done in such fields as population and demography, sociology of organization, and general sociology. The Tavistock Institute of Human Relations in London has become world famous and concentrates on human relations in the family, the work group, and organizations.

A parallel growth took place in Canada, Australia, and New Zealand. Canada, with some apparent reluctance, allowed itself to be much influenced by American sociology and has built many new departments with sociologists trained in the United States.

The Scandinavian countries have also to a considerable extent adopted the methods and some of the content of American sociology, and the subject has had rapid development in many of the universities and in research institutes, some of which are connected with universities. There is also a considerable amount of interchange between sociologists in these countries.

Japan has a record of much sociological activity dating back to the 1870s. The Japanese Sociological Society (Nippon Shakai Gakkai), headquartered at the University of Tokyo, was founded in 1923; by 1960 there were about 150 universities and colleges with courses in the subject. In the early period sociology was nearly all imported; Comte and Spencer, and later Giddings and Gabriel Tarde, were their important theorists. After World War II there were rapid changes in sociology in Japan, with empirical research methods largely replacing the earlier philosophical style. Importations from American sociology became abundant.

Popular among these were industrial sociology, educational sociology, public opinion research, and the study of mass communications.

Sociology in the former Soviet Union was long held back by the perceived incompatibility of the subject with Marxist theory. Eventually, however, it was permitted to develop, and sociological institutes and chairs of sociology increased. By 1970 the Soviet Sociological Association had more than a thousand members. Leading research interests included such subjects as labour productivity, education, crime, and alcoholism. Soviet sociology generally displayed an apparent tendency to avoid issues that might have implied conflict with Marxist thought.

Nations under the influence of the Soviet Union were also from time to time inhospitable to sociology, but the strong interest of younger scholars made possible some relaxation of this opposition, and in the second half of the 20th century there was considerable progress of sociology in such countries as Hungary, Poland, the Czech Republic, and Slovakia, with occasional setbacks in some areas.

In Israel the dominant department of sociology is at the Hebrew University in Jerusalem, where there are also several research institutes. Israeli sociology maintains continuous close contacts with American sociology, and many of the leading Israeli sociologists have had training or teaching experience in the United States. Among the specialties in Israel are research in methodology, communication, criminology, and the collective settlements (kibbutzim) in which new forms of custom and social organization are observed while under development.

The passing of the Fascist regime in Italy and the relative liberalization in Spain opened the door to sociology, and academic chairs and research institutes are gradually increasing in these countries. Of particular interest are studies of industrial efficiency and social mobility. The general conservatism of universities, however, may constitute a retarding influence for some time to come.

In Latin America objective sociology has been much resisted, partly because it has been viewed as a threat to the political and social order but also because of meagre financial support of research and the low salary level of professors, many of whom

must supplement their earnings in the practice of law, in civil service, and in other occupations. In the 1960s, however, the number of full-time chairs increased, and a number of research institutes, some financed by U.S. funds, were established. Political instability in some countries remains a major hindrance, and in such countries able scholars continue to be forced from their university positions from time to time.

Little by little, sociology is penetrating into some of the developing nations. A number of African universities have formed departments, and the subject is gaining in importance in the Philippines, India, Indonesia, and Pakistan.

Scientific status.

It is evident that sociology has not achieved triumphs comparable to those of the several older and more heavily supported sciences. A variety of interpretations have been offered to explain the difference - most frequently, that the growth of knowledge in the science of sociology is more random than cumulative. The true situation appears to be that in some parts of the discipline - such as methodology, ecology, demography, the study of social differentiation and mobility, attitude research, and the study of small-group interaction processes, public opinion, and mass communication - there has in fact taken place a slow but accelerating accumulation of organized and tested knowledge. In some other fields the expansion of the volume of literature has not appeared to have had this property. Critics have attributed the slow pace to a variety of factors - the appetite of sociologists for neologisms and jargon, a disposition for pseudoquantification, and excessive concern with imitation of the methods of natural sciences, overdependence on data from interviews, questionnaires, and informal observations. All these shortcomings can be found in contemporary sociology, but none is characteristic of all areas. In general there has been progress toward efficient terminology and methods and toward more satisfactory data, and conclusions are increasingly based on the harmonious mixture of research methods

applied to varied and repeated studies, and therefore are less dependent on the strength of one particular methodological device.

Bias, in more than one direction, is sometimes presumed to be a chronic affliction of sociology. This may arise in part from the fact that the subject matter of sociology is familiar and important in the daily life of everyone, so that there exist many opportunities for the abundant variations in philosophical outlook and individual preferences to appear as irrational bias. Thus critics have expressed disapproval of the sociologists' skepticism on various matters of faith, of their amoral relativism concerning customs, of their apparent oversimplifications of some principles, and of their particular fashions in categorization and abstraction. But skepticism toward much of the content of folk knowledge is a characteristic of all science, and relativism can be interpreted as merely an avoidance of antiscientific ethnocentrism.

Furthermore, abstraction, categorization, and simplification are necessary to the advancement of knowledge, and no one system satisfies everyone.

The dispute about the main purpose of sociology, whether it works to understand behaviour, or to cause social change, is a dispute found in every pursuit of scientific knowledge, and such polarization is far from absolute. Persons differ in the degree to which they regard the value of science as an intellectual understanding of the cosmos or as an instrument for immediate improvement of the human lot. Since even the "purest" scientist conceives of his work as benefiting mankind, the issue narrows to a difference in preference between an ad hoc attack on immediate human problems and a long-run trust that basic knowledge, gathered without reference to present urgencies, is even more valuable. Sociologists differ on this issue; in some countries there is much pressure toward early practicality of results; in others, including the United States, the larger number of scholars and the principal sociological associations have shown preference for "basic science." In very recent times, however, there has emerged a radical movement among students in various countries involving advocacy of complete commitment to action on current political and social problems.

A degree of polarization has also arisen over the proper strategy for research - whether research should take its directions from the needs of society and mankind or from the evolving theoretical corpus of sociology. In nations that allow academic freedom such disputes are usually of low intensity, because each scholar selects his research interests on any basis he prefers, including that of personal taste. In this way presumably the motivation of the investigator is maximized.

Sociologists most interested in action express impatience at the claims of others who prefer to separate their research from personal values. Much of the dispute prevails only because the two sides argue past each other. There can be wide agreement that no human being is without personal values, that research forced to confirm a particular set of values is not good science, and that there can be scientific issues toward which a particular investigator is value-neutral. In research that is susceptible to contamination by the values of the worker, it is generally possible to minimize the damage by employing methodological devices that help to insulate the scientist from his wishes for a particular outcome - such devices as objective observational techniques and measurement methods, independent and blind analysis of results, and so forth.

Current trends.

It would appear that the growth of sociology will accelerate in the visible future. Among present trends suggesting this likelihood are the increase in public appreciation of the subject, the expansion of available funds for both teaching and research, the steady reduction of sectarian opposition to inquiry into social institutions, the improvement in research methods and methods for gathering data that qualify for modern statistical treatment, and the growth of acceptance and support from scientists in other fields. There are possible factors that could inhibit such growth, such as some forms of extreme nationalism and internal conflict, but such conditions so far have impeded development only locally and temporarily.

Furthermore, it appears likely that public interest in the development of sociological knowledge will increase as a consequence of rising awareness of its

promise for human safety and welfare. As the expansion of civilization, with its advanced science and technology, progressively conquers the natural hazards that afflict preliterate and preindustrial peoples and diminishes such threats as natural catastrophes, famine, and disease, a wide range of new problems emerges. These are not the menaces of an impersonal nature, but dangers that arise from imperfection in human behaviour, particularly in organized human relations. Wars have shown a tendency to become larger and ever more destructive, and the causes, though far from being understood, clearly lie, in large measure, in the complexities of social organization, in the interaction of great corporate national bodies. There appears to be little hope that politics, unaided by social science among other disciplines, could reverse this trend.

Domestic problems within nations, regions, cities, and towns appear also to become increasing sources of human troubles. There is a general rise in the severity of ethnic hostilities, and of internal conflicts between generations, political factions, and other divisions of the populations. There are also threats to human welfare from various forms of general social disorganization, reflected in the spread of pockets of poverty, crime, vice, political corruption, and family disorganization. In recent times the threats of overpopulation and potential destruction of the ecological environment have added a further reason for public alarm. Contemporary sociology obviously does not yet provide the solutions, but what prospects of human survival there are depend a great deal on the increase of the applicable knowledge of various social sciences, including sociology.

Emerging subfields.

Because human behaviour observes no limits in its directions, it is possible for sociologists to extend their inquiries accordingly. The expansion of sociological interests thus has involved some penetration of adjacent traditional academic fields, such as political science, economics, anthropology, psychology, communications, speech, and to some extent even physiology and zoology. Fields within traditional sociology have also broadened their content, producing such

expanded subjects as ecology and comparative sociology. Not all this extension is new, however, since much of the 19th century sociology was also very broad, especially the cosmic sociology of one worker, Lester F. Ward, who conceived sociology as the science of sciences, properly covering and organizing all knowledge.

Applications of sociology also appear to be spreading in a variety of directions, and here the possibilities seem unlimited. Sociologists aid industries in obtaining more efficient production; they help unions to increase their power; they organize rebellions of young persons, reform disorganized villages, counsel persons and families, and give or sell services to a wide variety of consumers. To what extent these applied activities will continue to spread will doubtless depend on their effectiveness relative to other means of gaining the same effects.

There is also an expansion of sociology into other than practical applications; for example, there is mathematical sociology, in which mathematical models of social behaviour are developed without systematic observations of behaviour. These efforts are not directed toward immediate human use, but may have value as bases for comparison with real behaviour and thus aid explanation of behavioral causes. A mathematical model of a completely just theoretical process of social mobility, for example, could be useful as a standard for comparing actual mobility at different times and in different nations.

Emerging methodologies.

After the easiest sociological questions have been answered, the further progress of research requires ever greater effort and cost, and the proportion of discoveries by individual investigators declines as the necessity for larger teamwork research expands. This foreshadows increasing complexity of the organization of research, as has already taken place in older sciences. Large-scale research in sociology is made possible, and perhaps inevitable, by the availability of expensive computers, elaborate techniques of multivariate analysis, and the storage of information in the form of data banks and the like.

The strongest methodological emphasis in the near future is likely to be on the processes of rigorous testing of generalizations that now appear to be of strategic value in the general structure of sociological knowledge. Complete surprises in the field of human behaviour are less likely than in other sciences, since most of the possible human situations have been familiar in folk knowledge as well as in academic sociology. But the subject contains many inconsistent principles, and few of these have been put to a definitive test, partly from lack of adequate methodology and to some extent from shortage of funds and scientific manpower.

Emerging roles for sociologists.

In general the principal employment of sociologists has been in educational institutions, but recently, in various countries, there has been an increasing penetration into other fields of activity. Sociologists, particularly in earlier decades, have been involved in various organized agencies devoted to social work. They also have participated in government work at various levels, from the lower bureaucratic ranks all the way to high administrative responsibility, and in the case of Thomás Masaryk, former president of Czechoslovakia, to the highest office of a nation. In the United States sociologists have been extensively employed in the Bureau of the Census; the Bureau of the Budget; the Institutes of Health; various other sections of the Department of Health, Education, and Welfare; and the office of the president, where they have made contributions to policy.

Other directions of sociological activity include the roles of consultant, social critic, social activist, and even revolutionary. When the activity diverges far enough from traditional academic sociology, it may cease to be regarded as sociological, but it appears likely that sociologists will continue to spread their activities over the ever-widening region of national or global concern, in the name of their science or otherwise.

ФАКУЛЬТЕТ МАТЕМАТИКИ И ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ

An Introduction to Artificial Intelligence.

Artificial Intelligence, or Al for short, is a combination of computer science, physiology, and philosophy. Al is a broad topic, consisting of different fields, from machine vision to expert systems. The element that the fields of Al have in common is the creation of machines that can "think".

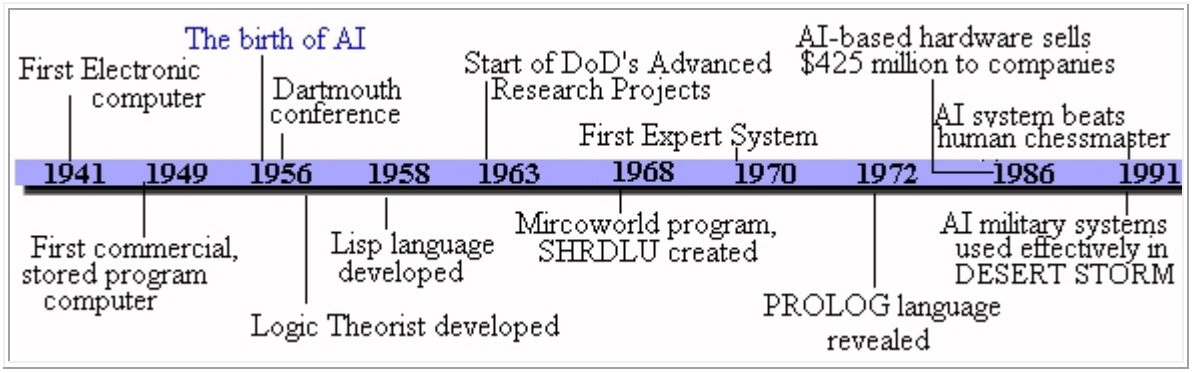
In order to classify machines as "thinking", it is necessary to define intelligence. To what degree does intelligence consist of, for example, solving complex problems, or making generalizations and relationships? And what about perception and comprehension? Research into the areas of learning, of language, and of sensory perception have aided scientists in building intelligent machines. One of the most challenging approaches facing experts is building systems that mimic the behavior of the human brain, made up of billions of neurons, and arguably the most complex matter in the universe. Perhaps the best way to gauge the intelligence of a machine is British computer scientist Alan Turing's test. He stated that a computer would deserves to be called intelligent if it could deceive a human into believing that it was human.

Artificial Intelligence has come a long way from its early roots, driven by dedicated researchers. The beginnings of Al reach back before electronics, to philosophers and mathematicians such as Boole and others theorizing on principles that were used as the foundation of Al Logic. Al really began to intrigue researchers with the invention of the computer in 1943. The technology was finally available, or so it seemed, to simulate intelligent behavior. Over the next four decades, despite many stumbling blocks, Al has grown from a dozen researchers, to thousands of engineers and specialists; and from programs capable of playing checkers, to systems designed to diagnose disease.

Al has always been on the pioneering end of computer science. Advanced-level computer languages, as well as computer interfaces and word-processors owe their existence to the research into artificial intelligence. The theory and insights

brought about by Al research will set the trend in the future of computing. The products available today are only bits and pieces of what are soon to follow, but they are a movement towards the future of artificial intelligence. The advancements in the quest for artificial intelligence have, and will continue to affect our jobs, our education, and our lives.

The History of Artificial Intelligence



Timeline of major AI events

Evidence of Artificial Intelligence folklore can be traced back to ancient Egypt, but with the development of the electronic computer in 1941, the technology finally became available to create machine intelligence. The term artificial intelligence was first coined in 1956, at the Dartmouth conference, and since then Artificial Intelligence has expanded because of the theories and principles developed by its dedicated researchers. Through its short modern history, advancement in the fields of AI have been slower than first estimated, progress continues to be made. From its birth 4 decades ago, there have been a variety of AI programs, and they have impacted other technological advancements.

The Era of the Computer:

In 1941 an invention revolutionized every aspect of the storage and processing of information. That invention, developed in both the US and Germany was the electronic computer. The first computers required large, separate air-conditioned rooms, and were a programmers nightmare, involving the separate configuration of thousands of wires to even get a program running.

The 1949 innovation, the stored program computer, made the job of entering a program easier, and advancements in computer theory lead to computer science, and eventually Artificial intelligence. With the invention of an electronic means of processing data, came a medium that made AI possible.

The Beginnings of AI:

Although the computer provided the technology necessary for AI, it was not until the early 1950's that the link between human intelligence and machines was really observed. Norbert Wiener was one of the first Americans to make observations on the principle of feedback theory feedback theory. The most familiar example of feedback theory is the thermostat: It controls the temperature of an environment by gathering the actual temperature of the house, comparing it to the desired temperature, and responding by turning the heat up or down. What was so important about his research into feedback loops was that Wiener theorized that all intelligent behavior was the result of feedback mechanisms. Mechanisms that could possibly be simulated by machines. This discovery influenced much of early development of AI.

In late 1955, Newell and Simon developed The Logic Theorist, considered by many to be the first AI program. The program, representing each problem as a tree model, would attempt to solve it by selecting the branch that would most likely result in the correct conclusion. The impact that the logic theorist made on both the public and the field of AI has made it a crucial stepping stone in developing the AI field.

In 1956 John McCarthy regarded as the father of AI, organized a conference to draw the talent and expertise of others interested in machine intelligence for a month of brainstorming. He invited them to Vermont for "The Dartmouth summer research project on artificial intelligence." From that point on, because of McCarthy, the field would be known as Artificial intelligence. Although not a huge success, (explain) the Dartmouth conference did bring together the founders in AI, and served to lay the groundwork for the future of AI research.

Knowledge Expansion

In the seven years after the conference, AI began to pick up momentum. Although the field was still undefined, ideas formed at the conference were re-examined, and built upon. Centers for AI research began forming at Carnegie Mellon and MIT, and a new challenges were faced: further research was placed upon creating systems that could efficiently solve problems, by limiting the search, such as the Logic Theorist. And second, making systems that could learn by themselves.

In 1957, the first version of a new program The General Problem Solver(GPS) was tested. The program developed by the same pair which developed the Logic Theorist. The GPS was an extension of Wiener's feedback principle, and was capable of solving a greater extent of common sense problems. A couple of years after the GPS, IBM contracted a team to research artificial intelligence. Herbert Gelerneter spent 3 years working on a program for solving geometry theorems.

While more programs were being produced, McCarthy was busy developing a major breakthrough in AI history. In 1958 McCarthy announced his new development; the LISP language, which is still used today. LISP stands for LISt Processing, and was soon adopted as the language of choice among most AI developers.

In 1963 MIT received a 2.2 million dollar grant from the United States government to be used in researching Machine-Aided Cognition (artificial intelligence). The grant by the Department of Defense's Advanced research projects Agency (ARPA), to ensure that the US would stay ahead of the Soviet Union in technological advancements. The project served to increase the pace of development in AI research, by drawing computer scientists from around the world, and continues funding.

The Multitude of programs

The next few years showed a multitude of programs, one notably was SHRDLU. SHRDLU was part of the microworlds project, which consisted of research and programming in small worlds (such as with a limited number of geometric shapes). The MIT researchers headed by Marvin Minsky, demonstrated that when confined to a small subject matter, computer programs could solve spatial problems and

logic problems. Other programs which appeared during the late 1960's were STUDENT, which could solve algebra story problems, and SIR which could understand simple English sentences. The result of these programs was a refinement in language comprehension and logic.

Another advancement in the 1970's was the advent of the expert system. Expert systems predict the probability of a solution under set conditions. For example:

Because of the large storage capacity of computers at the time, expert systems had the potential to interpret statistics, to formulate rules. And the applications in the market place were extensive, and over the course of ten years, expert systems had been introduced to forecast the stock market, aiding doctors with the ability to diagnose disease, and instruct miners to promising mineral locations. This was made possible because of the systems ability to store conditional rules, and a storage of information.

During the 1970's Many new methods in the development of AI were tested, notably Minsky's frames theory. Also David Marr proposed new theories about machine vision, for example, how it would be possible to distinguish an image based on the shading of an image, basic information on shapes, color, edges, and texture. With analysis of this information, frames of what an image might be could then be referenced. another development during this time was the PROLOGUE language. The language was proposed for In 1972, During the 1980's AI was moving at a faster pace, and further into the corporate sector. In 1986, US sales of AI-related hardware and software surged to $425 million. Expert systems in particular demand because of their efficiency. Companies such as Digital Electronics were using XCON, an expert system designed to program the large VAX computers. DuPont, General Motors, and Boeing relied heavily on expert systems Indeed to keep up with the demand for the computer experts, companies such as Teknowledge and Intellicorp specializing in creating software to aid in producing expert systems formed. Other expert systems were designed to find and correct flaws in existing expert systems.

The Transition from Lab to Life

The impact of the computer technology, AI included was felt. No longer was the computer technology just part of a select few researchers in laboratories. The personal computer made its debut along with many technological magazines. Such foundations as the American Association for Artificial Intelligence also started. There was also, with the demand for AI development, a push for researchers to join private companies. 150 companies such as DEC which employed its AI research group of 700 personnel, spend $1 billion on internal AI groups.

Other fields of AI also made there way into the marketplace during the 1980's. One in particular was the machine vision field. The work by Minsky and Marr were now the foundation for the cameras and computers on assembly lines, performing quality control. Although crude, these systems could distinguish differences shapes in objects using black and white differences. By 1985 over a hundred companies offered machine vision systems in the US, and sales totaled $80 million.

The 1980's were not totally good for the AI industry. In 1986-87 the demand in AI systems decreased, and the industry lost almost a half of a billion dollars. Companies such as Teknowledge and Intellicorp together lost more than $6 million, about a third of there total earnings. The large losses convinced many research leaders to cut back funding. Another disappointment was the so called "smart truck" financed by the Defense Advanced Research Projects Agency. The projects goal was to develop a robot that could perform many battlefield tasks. In 1989, due to project setbacks and unlikely success, the Pentagon cut funding for the project.

Despite these discouraging events, AI slowly recovered. New technology in Japan was being developed. Fuzzy logic, first pioneered in the US has the unique ability to make decisions under uncertain conditions. Also neural networks were being reconsidered as possible ways of achieving Artificial Intelligence. The 1980's introduced to its place in the corporate marketplace, and showed the technology had real life uses, ensuring it would be a key in the 21st century.

AI put to the Test

The military put AI based hardware to the test of war during Desert Storm. AI-based technologies were used in missile systems, heads-up-displays, and other advancements. AI has also made the transition to the home. With the popularity of the AI computer growing, the interest of the public has also grown. Applications for the Apple Macintosh and IBM compatible computer, such as voice and character recognition have become available. Also AI technology has made steadying camcorders simple using fuzzy logic. With a greater demand for AI-related technology, new advancements are becoming available. Inevitably Artificial Intelligence has, and will continue to affecting our lives.

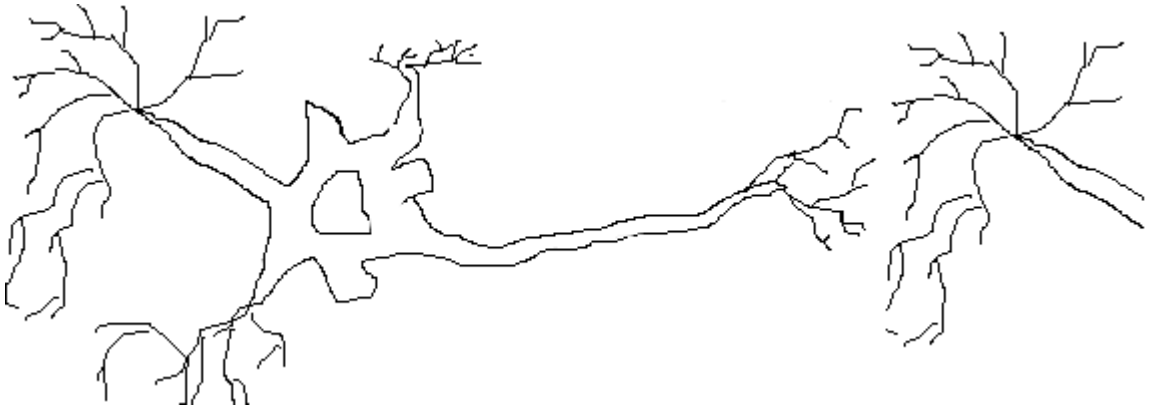
Methods used to create intelligence.

Introduction

In the quest to create intelligent machines, the field of Artificial Intelligence has split into several different approaches based on the opinions about the most promising methods and theories. These rivaling theories have lead researchers in one of two basic approaches; bottom-up and top-down. Bottom-up theorists believe the best way to achieve artificial intelligence is to build electronic replicas of the human brain's complex network of neurons, while the top-down approach attempts to mimic the brain's behavior with computer programs.

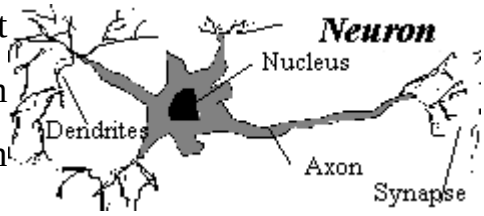
Neural Networks and Parallel Computation

The human brain is made up of a web of billions of cells called neurons, and understanding its complexities is seen as one of the last frontiers in scientific research. It is the aim of AI researchers who prefer this bottom-up approach to construct electronic circuits that act as neurons do in the human brain. Although much of the working of the brain remains unknown, the complex network of neurons is what gives humans intelligent characteristics. By itself, a neuron is not intelligent, but when grouped together, neurons are able to pass electrical signals through networks.



The neuron "firing", passing a signal to the next in the chain.

Research has shown that a signal received by a neuron travels through the dendrite region, and down the axon. Separating nerve cells is a gap called the synapse. In order for the signal to be transferred to the next neuron, the signal must be converted from electrical to chemical energy. The signal can then be received by the next neuron and processed.



Warren McCulloch after completing medical school at Yale, along with Walter Pitts a mathematician proposed a hypothesis to explain the fundamentals of how neural networks made the brain work. Based on experiments with neurons, McCulloch and Pitts showed that neurons might be considered devices for processing binary numbers. An important back of mathematic logic, binary numbers (represented as 1's and 0's or true and false) were also the basis of the electronic computer. This link is the basis of computer-simulated neural networks, also know as Parallel computing.

A century earlier the true / false nature of binary numbers was theorized in 1854 by George Boole in his postulates concerning the Laws of Thought. Boole's principles make up what is known as Boolean algebra, the collection of logic concerning AND, OR, NOT operands. For example according to the Laws of thought the statement: (for this example consider all apples red)

Apples are red - is True

Apples are red AND oranges are purple - is False

Apples are red OR oranges are purple - is True

Apples are red AND oranges are NOT purple - is also True

Boole also assumed that the human mind works according to these laws, it performs logical operations that could be reasoned. Ninety years later, Claude Shannon applied Boole's principles in circuits, the blueprint for electronic computers. Boole's contribution to the future of computing and Artificial Intelligence was immeasurable, and his logic is the basis of neural networks.

McCulloch and Pitts, using Boole's principles, wrote a paper on neural network theory. The thesis dealt with how the networks of connected neurons could perform logical operations. It also stated that, one the level of a single neuron, the release or failure to release an impulse was the basis by which the brain makes true / false decisions. Using the idea of feedback theory, they described the loop which existed between the senses - -> brain - -> muscles, and likewise concluded that Memory could be defined as the signals in a closed loop of neurons. Although we now know that logic in the brain occurs at a level higher then McCulloch and Pitts theorized, their contributions were important to AI because they showed how the firing of signals between connected neurons could cause the brains to make decisions. McCulloch and Pitt's theory is the basis of the artificial neural network theory.

Using this theory, McCulloch and Pitts then designed electronic replicas of neural networks, to show how electronic networks could generate logical processes. They also stated that neural networks may, in the future, be able to learn, and recognize patterns. The results of their research and two of Weiner's books served to increase enthusiasm, and laboratories of computer simulated neurons were set up across the country.

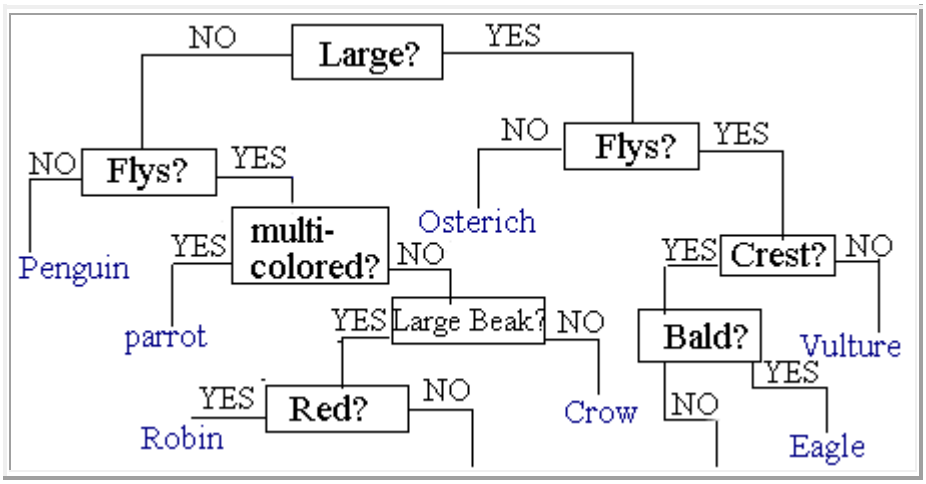
Two major factors have inhibited the development of full scale neural networks. Because of the expense of constructing a machine to simulate neurons, it was expensive even to construct neural networks with the number of neurons in an ant. Although the cost of components have decreased, the computer would have to grow thousands of times larger to be on the scale of the human brain. The second factor is current computer architecture. The standard Von Neuman computer, the

architecture of nearly all computers, lacks an adequate number of pathways between components. Researchers are now developing alternate architectures for use with neural networks.

Even with these inhibiting factors, artificial neural networks have presented some impressive results. Frank Rosenblatt, experimenting with computer simulated networks, was able to create a machine that could mimic the human thinking process, and recognize letters. But, with new top-down methods becoming popular, parallel computing was put on hold. Now neural networks are making a return, and some researchers believe that with new computer architectures, parallel computing and the bottom-up theory will be a driving factor in creating artificial intelligence.

Top Down Approaches; Expert Systems

Because of the large storage capacity of computers, expert systems had the potential to interpret statistics, in order to formulate rules. An expert system works much like a detective solves a mystery. Using the information, and logic or rules, an expert system can solve the problem. For example it the expert system was designed to distinguish birds it may have the following:



Charts like these represent the logic of expert systems. Using a similar set of rules, experts can have a variety of applications. With improved interfacing, computers may begin to find a larger place in society.

Chess

AI-based game playing programs combine intelligence with entertainment. On game with strong AI ties is chess. World-champion chess playing programs can see

ahead twenty plus moves in advance for each move they make. In addition, the programs have an ability to get progressably better over time because of the ability to learn. Chess programs do not play chess as humans do. In three minutes, Deep Thought (a master program) considers 126 million moves, while human chessmaster on average considers less than 2 moves. Herbert Simon suggested that human chess masters are familiar with favorable board positions, and the relationship with thousands of pieces in small areas. Computers on the other hand, do not take hunches into account. The next move comes from exhaustive searches into all moves, and the consequences of the moves based on prior learning. Chess programs, running on Cray super computers have attained a rating of 2600 (senior master), in the range of Gary Kasparov, the Russian world champion.

Frames

On method that many programs use to represent knowledge are frames. Pioneered by Marvin Minsky, frame theory revolves around packets of information. For example, say the situation was a birthday party. A computer could call on its birthday frame, and use the information contained in the frame, to apply to the situation. The computer knows that there is usually cake and presents because of the information contained in the knowledge frame. Frames can also overlap, or contain sub-frames. The use of frames also allows the computer to add knowledge. Although not embraced by all AI developers, frames have been used in comprehension programs such as Sam.

Conclusion

This page touched on some of the main methods used to create intelligence. These approaches have been applied to a variety of programs. As we progress in the development of Artificial Intelligence, other theories will be available, in addition to building on today's methods.

Essays on the use of AI.

What we can do with AI

We have been studying this issue of AI application for quite some time now and know all the terms and facts. But what we all really need to know is what can we do to get our hands on some AI today. How can we as individuals use our own technology? We hope to discuss this in depth (but as briefly as possible) so that you the consumer can use AI as it is intended.

First, we should be prepared for a change. Our conservative ways stand in the way of progress. AI is a new step that is very helpful to the society. Machines can do jobs that require detailed instructions followed and mental alertness. AI with its learning capabilities can accomplish those tasks but only if the worlds conservatives are ready to change and allow this to be a possibility. It makes us think about how early man finally accepted the wheel as a good invention, not something taking away from its heritage or tradition.

Secondly, we must be prepared to learn about the capabilities of AI. The more use we get out of the machines the less work is required by us. In turn less injuries and stress to human beings. Human beings are a species that learn by trying, and we must be prepared to give AI a chance seeing AI as a blessing, not an inhibition.

Finally, we need to be prepared for the worst of AI. Something as revolutionary as AI is sure to have many kinks to work out. There is always that fear that if AI is learning based, will machines learn that being rich and successful is a good thing, then wage war against economic powers and famous people? There are so many things that can go wrong with a new system so we must be as prepared as we can be for this new technology.

However, even though the fear of the machines are there, their capabilities are infinite Whatever we teach AI, they will suggest in the future if a positive outcome arrives from it. AI are like children that need to be taught to be kind, well mannered, and intelligent. If they are to make important decisions, they should be wise. We as citizens need to make sure AI programmers are keeping things on the level. We should be sure they are doing the job correctly, so that no future accidents occur.

AIAI Teaching Computers Computers

Does this sound a little Redundant? Or maybe a little redundant? Well just sit back and let me explain. The Artificial Intelligence Applications Institute has many project that they are working on to make their computers learn how to operate themselves with less human input. To have more functionality with less input is an operation for AI technology. I will discuss just two of these projects: AUSDA and EGRESS.

AUSDA is a program which will exam software to see if it is capable of handling the tasks you need performed. If it isn't able or isn't reliable AUSDA will instruct you on finding alternative software which would better suit your needs. According to AIAI, the software will try to provide solutions to problems like "identifying the root causes of incidents in which the use of computer software is involved, studying different software development approaches, and identifying aspects of these which are relevant to those root causes producing guidelines for using and improving the development approaches studied, and providing support in the integration of these approaches, so that they can be better used for the development and maintenance of safety critical software."

Sure, for the computer buffs this program is a definitely good news. But what about the average person who think the mouse is just the computers foot pedal? Where do they fit into computer technology. Well don't worry guys, because us nerds are looking out for you too! Just ask AIAI what they have for you and it turns up the EGRESS is right down your alley. This is a program which is studying human reactions to accidents. It is trying to make a model of how peoples reactions in panic moments save lives. Although it seems like in tough situations humans would fall apart and have no idea what to do, it is in fact the opposite. Quick Decisions are usually made and are effective but not flawless. These computer models will help rescuers make smart decisions in time of need. AI can't be positive all the time but can suggest actions which we can act out and therefor lead to safe rescues.

So AIAI is teaching computers to be better computers and better people. AI technology will never replace man but can be an extension of our body which

allows us to make more rational decisions faster. And with Institutes like AIAI- we continue each stay to step forward into progress.

No worms in these Apples by Adam Dyess

Apple Computers may not have ever been considered as the state of art in Artificial Intelligence, but a second look should be given. Not only are today's PC's becoming more powerful but AI influence is showing up in them. From Macros to Voice Recognition technology, PC's are becoming our talking buddies. Who else would go surfing with you on short notice- even if it is the net. Who else would care to tell you that you have a business appointment scheduled at 8:35 and 28 seconds and would notify you about it every minute till you told it to shut up. Even with all the abuse we give today's PC's they still plug away to make us happy. We use PC's more not because they do more or are faster but because they are getting so much easier to use. And their ease of use comes from their use of AI.

All Power Macintoshes come with Speech Recognition. That's right- you tell the computer to do what you want without it having to learn your voice. This implication of AI in Personal computers is still very crude but it does work given the correct conditions to work in and a clear voice. Not to mention the requirement of at least 16Mgs of RAM for quick use. Also Apple's Newton and other hand held note pads have Script recognition. Cursive or Print can be recognized by these notepad sized devices. With the pen that accompanies your silicon note pad you can write a little note to yourself which magically changes into computer text if desired. No more complaining about sloppy written reports if your computer can read your handwriting. If it can't read it though- perhaps in the future, you can correct it by dictating your letters instead.

Macros provide a huge stress relief as your computer does faster what you could do more tediously. Macros are old but they are to an extent, Intelligent. You have taught the computer to do something only by doing it once. In businesses, many times applications are upgraded. But the files must be converted. All of the businesses records but be changed into the new software's type. Macros save the work of conversion of hundred of files by a human by teaching the computer to

mimic the actions of the programmer. Thus teaching the computer a task that it can repeat whenever ordered to do so.

AI is all around us all but get ready for a change. But don't think the change will be harder on us because AI has been developed to make our lives easier.

The Scope of Expert Systems

As stated in the 'approaches' section, an expert system is able to do the work of a professional. Moreover, a computer system can be trained quickly, has virtually no operating cost, never forgets what it learns, never calls in sick, retires, or goes on vacation. Beyond those, intelligent computers can consider a large amount of information that may not be considered by humans.

But to what extent should these systems replace human experts? Or, should they at all? For example, some people once considered an intelligent computer as a possible substitute for human control over nuclear weapons, citing that a computer could respond more quickly to a threat. And many AI developers were afraid of the possibility of programs like Eliza, the psychiatrist and the bond that humans were making with the computer. We cannot, however, over look the benefits of having a computer expert. Forecasting the weather, for example, relies on many variables, and a computer expert can more accurately pool all of its knowledge. Still a computer cannot rely on the hunches of a human expert, which are sometimes necessary in predicting an outcome.

In conclusion, in some fields such as forecasting weather or finding bugs in computer software, expert systems are sometimes more accurate than humans. But for other fields, such as medicine, computers aiding doctors will be beneficial, but the human doctor should not be replaced. Expert systems have the power and range to aid to benefit, and in some cases replace humans, and computer experts, if used with discretion, will benefit human kind.

Information Processing and Information Systems

Introduction

In popular usage, the term information refers to facts and opinions provided and received during the course of daily life: one obtains information directly from other living beings, from mass media, from electronic data banks, and from all sorts of observable phenomena in the surrounding environment. A person using such facts and opinions generates more information, some of which is communicated to others during discourse, by instructions, in letters and documents, and through other media. Information organized according to some logical relationships is referred to as a body of knowledge, to be acquired by systematic exposure or study. Application of knowledge (or skills) yields expertise, and additional analytical or experiential insights are said to constitute instances of wisdom. Use of the term information is not restricted exclusively to its communication via natural language. Information is also registered and communicated through art and by facial expressions and gestures or by such other physical responses as shivering. Moreover, every living entity is endowed with information in the form of a genetic code. These information phenomena permeate the physical and mental world, and their variety is such that it has defied so far all attempts at a unified definition of information.

Interest in information phenomena has increased dramatically in the 20th century, and today they are the objects of study in a number of disciplines, including philosophy, physics, biology, linguistics, information and computer science, electronic and communications engineering, management science, and the social sciences. On the commercial side, the information service industry has become one of the newer industries worldwide. Almost all other industries - manufacturing and service - are increasingly concerned with information and its handling. The different, though often overlapping, viewpoints and phenomena of these fields lead to different (and sometimes conflicting) concepts and "definitions" of information.

This article touches on such concepts, particularly as they relate to information processing and information systems. In treating the basic elements of information processing, it distinguishes between information in analog and digital form, and it describes their acquisition, recording, organization, retrieval, display, and dissemination. In treating information systems, the article discusses system analysis and design and provides a descriptive taxonomy of the main system types. Some attention is also given to the social impact of information systems and to the field of information science.

General considerations

BASIC CONCEPTS

Interest in how information is communicated and how its carriers convey meaning has occupied, since the time of pre-Socratic philosophers, the field of inquiry called semiotics, the study of signs and sign phenomena. Signs are the irreducible elements of communication and the carriers of meaning. The American philosopher, mathematician, and physicist Charles S. Peirce is credited with having pointed out the three dimensions of signs, which are concerned with, respectively, the body or medium of the sign, the object that the sign designates, and the interpretant or interpretation of the sign. Peirce recognized that the fundamental relations of information are essentially triadic; in contrast, all relations of the physical sciences are reducible to dyadic (binary) relations. Another American philosopher, Charles W. Morris, designated these three sign dimensions syntactic, semantic, and pragmatic, the names by which they are known today.

Information processes are executed by information processors. For a given information processor, whether physical or biological, a token is an object, devoid of meaning, that the processor recognizes as being totally different from other tokens. A group of such unique tokens recognized by a processor constitutes its basic "alphabet"; for example, the dot, dash, and space constitute the basic token alphabet of a Morse-code processor. Objects that carry meaning are represented by patterns of tokens called symbols. The latter combine to form symbolic

expressions that constitute inputs to or outputs from information processes and are stored in the processor memory.

Information processors are components of an information system, which is a class of constructs. An abstract model of an information system features four basic elements: processor, memory, receptor, and effector. The processor has several functions: (1) to carry out elementary information processes on symbolic expressions, (2) to store temporarily in the processor's short-term memory the input and output expressions on which these processes operate and which they generate, (3) to schedule execution of these processes, and (4) to change this sequence of operations in accordance with the contents of the short-term memory. The memory stores symbolic expressions, including those that represent composite information processes, called programs. The two other components, the receptor and the effector, are input and output mechanisms whose functions are, respectively, to receive symbolic expressions or stimuli from the external environment for manipulation by the processor and to emit the processed structures back to the environment.

The power of this abstract model of an information-processing system is provided by the ability of its component processors to carry out a small number of elementary information processes: reading; comparing; creating, modifying, and naming; copying; storing; and writing. The model, which is representative of a broad variety of such systems, has been found useful to explicate man-made information systems implemented on sequential information processors.

Because it has been recognized that in nature information processes are not strictly sequential, increasing attention has been focused since 1980 on the study of the human brain as an information processor of the parallel type. The cognitive sciences, the interdisciplinary field that focuses on the study of the human mind, have contributed to the development of neurocomputers, a new class of parallel, distributed-information processors that mimic the functioning of the human brain, including its capabilities for self-organization and learning. So-called neural networks, which are mathematical models inspired by the neural circuit network of

the human brain, are increasingly finding applications in areas such as pattern recognition, control of industrial processes, and finance, as well as in many research disciplines.

INFORMATION AS A RESOURCE AND COMMODITY

In the late 20th century, information has acquired two major utilitarian connotations. On the one hand, it is considered an economic resource, somewhat on par with other resources such as labour, material, and capital. This view stems from evidence that the possession, manipulation, and use of information can increase the cost-effectiveness of many physical and cognitive processes. The rise in information-processing activities in industrial manufacturing as well as in human problem solving has been remarkable. Analysis of one of the three traditional divisions of the economy, the service sector, shows a sharp increase in information-intensive activities since the beginning of the 20th century. By 1975 these activities accounted for half of the labour force of the United States (see Table 1), giving rise to the so-called information society.

As an individual and societal resource, information has some interesting characteristics that separate it from the traditional notions of economic resources. Unlike other resources, information is expansive, with limits apparently imposed only by time and human cognitive capabilities. Its expansiveness is attributable to the following: (1) it is naturally diffusive; (2) it reproduces rather than being consumed through use; and (3) it can be shared only, not exchanged in transactions. At the same time, information is compressible, both syntactically and semantically. Coupled with its ability to be substituted for other economic resources, its transportability at very high speeds, and its ability to impart advantages to the holder of information, these characteristics are at the base of such societal industries as research, education, publishing, marketing, and even politics. Societal concern with the husbanding of information resources has extended from the traditional domain of libraries and archives to encompass

organizational, institutional, and governmental information under the umbrella of information resource management.

The second perception of information is that it is an economic commodity, which helps to stimulate the worldwide growth of a new segment of national economies - the information service sector. Taking advantage of the properties of information and building on the perception of its individual and societal utility and value, this sector provides a broad range of information products and services. By 1992 the market share of the U.S. information service sector had grown to about $25 billion (see Table 2). This was equivalent to about one-seventh of the country's computer market, which, in turn, represented roughly 40 percent of the global market in computers in that year. However, the probable convergence of computers and television (which constitutes a market share 100 times larger than computers) and its impact on information services, entertainment, and education are likely to restructure the respective market shares of the information industry before the onset of the 21st century.

Elements of information processing

Humans receive information with their senses: sounds through hearing; images and text through sight; shape, temperature, and affection through touch; and odours through smell. To interpret the signals received from the senses, humans have developed and learned complex systems of languages consisting of "alphabets" of symbols and stimuli and the associated rules of usage. This has enabled them to recognize the objects they see, understand the messages they read or hear, and comprehend the signs received through the tactile and olfactory senses.

The carriers of information-conveying signs received by the senses are energy phenomena - audio waves, light waves, and chemical and electrochemical stimuli. In engineering parlance, humans are receptors of analog signals; and, by a somewhat loose convention, the messages conveyed via these carriers are called analog-form information, or simply analog information. Until the development of the digital computer, cognitive information was stored and processed only in

analog form, basically through the technologies of printing, photography, and telephony.

Although humans are adept at processing information stored in their memories, analog information stored external to the mind is not processed easily. Modern information technology greatly facilitates the manipulation of externally stored information as a result of its representation as digital signals - i.e., as the presence or absence of energy (electricity, light, or magnetism). Information represented digitally in two-state, or binary, form is often referred to as digital information. Modern information systems are characterized by extensive metamorphoses of analog and digital information. With respect to information storage and communication, the transition from analog information to digital is so pervasive that the end of the 20th century will likely witness a historic transformation of the manner in which humans create, access, and use information.

ACQUISITION AND RECORDING OF INFORMATION IN ANALOG FORM

The principal categories of information sources useful in modern information systems are text, video, and voice. One of the first ways in which prehistoric humans communicated was by sound; sounds represented concepts such as pleasure, anger, and fear, as well as objects of the surrounding environment, including food and tools. Sounds assumed their meaning by convention - namely, by the use to which they were consistently put. Combining parts of sound allowed representation of more complex concepts, gradually leading to the development of speech and eventually to spoken "natural" languages.

For information to be communicated broadly, it needs to be stored external to human memory; accumulation of human experience, knowledge, and learning would be severely limited without such storage, making necessary the development of writing systems.

Civilization can be traced to the time when humans began to associate abstract shapes with concepts and with the sounds of speech that represented them. Early recorded representations were those of visually perceived objects and events, as, for example, the animals and activities depicted in Paleolithic cave drawings. The

evolution of writing systems proceeded through the early development of pictographic languages, in which a symbol would represent an entire concept. Such symbols would go through many metamorphoses of shape in which the resemblance between each symbol and the object it stood for gradually disappeared, but its semantic meaning would become more precise. As the conceptual world of humans became larger, the symbols, called ideographs, grew in number. Modern Chinese, a present-day result of this evolutionary direction of a pictographic writing system, has upward of 50,000 ideographs.

At some point in the evolution of written languages, the method of representation shifted from the pictographic to the phonetic: speech sounds began to be represented by an alphabet of graphic symbols. Combinations of a relatively small set of such symbols could stand for more complex concepts as words, phrases, and sentences. The invention of the written phonetic alphabet is thought to have taken place during the 2nd millennium BC. The pragmatic advantages of alphabetic writing systems over the pictographic became apparent twice in the present millennium: after the invention of the movable-type printing press in the 15th century and again with the development of information processing by electronic means since the mid-1940s.

From the time early humans learned to represent concepts symbolically, they used whatever materials were readily available in nature for recording. The Sumerian cuneiform, a wedge-shaped writing system, was impressed by a stylus into soft clay tablets, which were subsequently hardened by drying in the sun or the oven. The earliest Chinese writing, dating to the 2nd millennium BC, is preserved on animal bone and shell, while early writing in India was done on palm leaves and birch bark. Applications of technology yielded other materials for writing. The Chinese had recorded their pictographs on silk, using brushes made from animal hair, long before they invented paper. The Egyptians first wrote on cotton, but they began using papyrus sheets and rolls made from the fibrous lining of the papyrus plant during the 4th millennium BC. The reed brush and a palette of ink were the implements with which they wrote hieroglyphic script. Writing on parchment, a

material which was superior to papyrus and was made from the prepared skins of animals, became commonplace about 200 BC, some 300 years after its first recorded use, and the quill pen replaced the reed brush. By the 4th century AD, parchment came to be the principal writing material in Europe.

Paper was invented in China at the beginning of the 2nd century AD, and for some 600 years its use was confined to East Asia. In AD 751 Arab and Chinese armies clashed at the Battle of Talas, near Samarkand; among the Chinese taken captive were some papermakers from whom the Arabs learned the techniques. From the 7th century on, paper became the dominant writing material of the Islamic world. Papermaking finally reached Spain and Sicily in the 12th century, and it took another three centuries before it was practiced in Germany.

With the invention of printing from movable type, typesetting became the standard method of creating copy. Typesetting was an entirely manual operation until the adoption of a typewriter-like keyboard in the 19th century. In fact, it was the typewriter that mechanized the process of recording original text. Although the typewriter was invented during the early 18th century in England, the first practical version, constructed by the American inventor Christopher Latham Sholes, did not appear until 1867. The mechanical typewriter finally found wide use after World War I. Today its electronic variant, the computer video terminal, is used pervasively to record original text.

Recording of original nontextual (image) information was a manual process until the development of photography during the early decades of the 19th century; drawing and carving were the principal early means of recording graphics. Other techniques were developed alongside printing - for example, etching in stone and metal. The invention of film and the photographic process added a new dimension to information acquisition: for the first time, complex visual images of the real world could be captured accurately. Photography provided a method of storing information in less space and more accurately than was previously possible with narrative information.

During the 20th century, versatile electromagnetic media have opened up new possibilities for capturing original analog information. Magnetic audio tape is used to capture speech and music, and magnetic videotape provides a low-cost medium for recording analog voice and video signals directly and simultaneously. Magnetic technology has other uses in the direct recording of analog information, including alphanumerics. Magnetic characters, bar codes, and special marks are printed on checks, labels, and forms for subsequent sensing by magnetic or optical readers and conversion to digital form. Banks, educational institutions, and the retail industry rely heavily on this technology. Nonetheless, paper and film continue to be the dominant media for direct storage of textual and visual information in analog form.

ACQUISITION AND RECORDING OF INFORMATION IN DIGITAL FORM

The versatility of modern information systems stems from their ability to represent information electronically as digital signals and to manipulate it automatically at exceedingly high speeds. Information is stored in binary devices, which are the basic components of digital technology. Because these devices exist only in one of two states, information is represented in them either as the absence or the presence of energy (electric pulse). The two states of binary devices are conveniently designated by the binary digits, or bits, zero (0) and one (1).

In this manner, alphabetic symbols of natural-language writing systems can be represented digitally as combinations of zeros (no pulse) and ones (pulse). Tables of equivalences of alphanumeric characters and strings of binary digits are called coding systems, the counterpart of writing systems. A combination of three binary digits can represent up to eight such characters; one comprising four digits, up to 16 characters; and so on. The choice of a particular coding system depends on the size of the character set to be represented. The widely used systems are the American Standard Code for Information Interchange (ASCII), a seven- or eight-bit code representing the English alphabet, numerals, and certain special characters of the standard computer keyboard; and the corresponding eight-bit Extended Binary Coded Decimal Interchange Code (EBCDIC), used for computers produced

by IBM (International Business Machines Corp.) and most compatible systems. The digital representation of a character by eight bits is called a byte.

The seven-bit ASCII code is capable of representing up to 128 alphanumeric and special characters - sufficient to accommodate the writing systems of many phonetic scripts, including Latin and Cyrillic. Some alphabetic scripts require more than seven bits; for example, the Arabic alphabet, also used in the Urdu and Persian languages, has 28 consonantal characters (as well as a number of vowels and diacritical marks), but each of these may have four shapes, depending on its position in the word.

For digital representation of nonalphabetic writing systems, even the eight-bit code accommodating 256 characters is inadequate. Some writing systems that use Chinese characters, for example, have more than 50,000 ideographs (the minimal standard font for the Hanzi system in Chinese and the kanji system in Japanese has about 7,000 ideographs). Digital representation of such scripts can be accomplished in three ways. One approach is to develop a phonetic character set; the Chinese Pinyin, the Korean Hangul, and the Japanese hiragana phonetic schemes all have alphabetic sets similar in number to the Latin alphabet. As the use of phonetic alphabets in Oriental cultures is not yet widespread, they may be converted to ideographic by means of a dictionary lookup (see Figure 2). A second technique is to decompose ideographs into a small number of elementary signs called strokes, the sum of which constitutes a shape-oriented, nonphonetic alphabet. The third approach is to use more than eight bits to encode the large numbers of ideographs; for instance, two bytes can represent uniquely more than 65,000 ideographs. Because the eight-bit ASCII code is inadequate for a number of writing systems, either because they are nonalphabetic or because their phonetic scripts possess large numbers of diacritical marks, the computer industry in 1991 began formulating a new international coding standard based on 16 bits.

Recording media.

Punched cards and perforated paper tape were once widely used to store data in binary form. Today they have been supplanted by media based on electromagnetic and electro-optic technologies except in a few special applications.

Present-day storage media are of two types: random- and serial-, or sequential-, access. In random-access media (such as primary memory) the time required to access a given piece of data is independent of its location, while in serial-access media the access time depends on the data's location and the position of the read-write head. The typical serial-access medium is magnetic tape. The storage density of magnetic tape has increased considerably over the years, mainly by increases in the number of tracks packed across the width of the tape.

While magnetic tape remains a popular choice in applications requiring low-cost auxiliary storage and data exchange, new tape variants have begun entering the market of the 1990s. Video recording tape has been adapted for digital storage, and digital audio tape (DAT) surpasses all tape storage devices in offering the highest areal data densities. DAT technology uses a helical-scan recording method in which both the tape and the recording head move simultaneously, allowing extremely high recording densities. A four-millimetre DAT tape cassette has a capacity of up to eight billion bytes (eight gigabytes). The capacity of this tape is expected to increase by an order of magnitude well before the year 2000.

Another type of magnetic storage medium, the magnetic disk, provides rapid, random access to data. This device, developed in 1962, consists of either an aluminum or plastic platen coated with a metallic material. Information is recorded on a disk by turning the charge of the read-write head on and off, which produces magnetic "dots" representing binary digits in circular tracks. A block of data on a given track can be accessed without having to pass over a large portion of its contents sequentially, as in the case of tape. Data-retrieval time is thus reduced dramatically. Hard disk drives built into personal computers and workstations have storage capacities of up to several gigabytes. Large computers using disk cartridges can provide virtually unlimited mass storage.

During the 1970s the floppy disk - a small, flexible disk - was introduced for use in personal computers and other microcomputer systems. Compared with the storage capacity of the conventional hard disk, that of such a "soft" diskette is low - under three million characters. This medium is used primarily for loading and backing up personal computers.

An entirely different kind of recording and storage medium, the optical disc, became available during the early 1980s. The optical disc makes use of laser technology: digital data are recorded by burning a series of microscopic holes, or pits, with a laser beam into thin metallic film on the surface of a 4 3/4-inch (12-centimetre) plastic disc. In this way, information from magnetic tape is encoded on a master disc; subsequently, the master is replicated by a process called stamping. In the read mode, low-intensity laser light is reflected off the disc surface and is "read" by light-sensitive diodes. The radiant energy received by the diodes varies according to the presence of the pits, and this input is digitized by the diode circuits. The digital signals are then converted to analog information on a video screen or in printout form.

Since the introduction of this technology, three main types of optical storage media have become available: (1) rewritable, (2) write-once read-many (WORM), and (3) compact disc read-only memory (CD-ROM). Rewritable discs are functionally equivalent to magnetic disks, although the former are slower. WORM discs are used as an archival storage medium to enter data once and retrieve it many times. CD-ROMs are the preferred medium for electronic distribution of digital libraries and software. To raise storage capacity, optical discs are arranged into "jukeboxes" holding as many as 10 million pages of text or more than one terabyte (one trillion bytes) of image data. The high storage capacities and random access of the magneto-optical, rewritable discs are particularly suited for storing multimedia information, in which text, image, and sound are combined.

Recording techniques.

Digitally stored information is commonly referred to as data, and its analog counterpart is called source data. Vast quantities of nondocument analog data are

collected, digitized, and compressed automatically by means of appropriate instruments in fields such as astronomy, environmental monitoring, scientific experimentation and modeling, and national security. The capture of information generated by humankind, in the form of packages of symbols called documents, is accomplished by manual and, increasingly, automatic techniques. Data are entered manually by striking the keys of a keyboard, touching a computer screen, or writing by hand on a digital tablet or its recent variant, the so-called pen computer. Manual data entry, a slow and error-prone process, is facilitated to a degree by special computer programs that include editing software, with which to insert formatting commands, verify spelling, and make text changes, and document-formatting software, with which to arrange and rearrange text and graphics flexibly on the output page.

It is estimated that 5 percent of all documents in the United States exist in digitized form and that two-thirds of the paper documents cannot be digitized by keyboard transcription because they contain drawings or still images and because such transcription would be highly uneconomic. Such documents are digitized economically by a process called document imaging.

Document imaging utilizes digital scanners to generate a digital representation of a document page. An image scanner divides the page into minute picture areas called pixels and produces an array of binary digits, each representing the brightness of a pixel. The resulting stream of bits is enhanced and compressed (to as little as 10 percent of the original volume) by a device called an image controller and is stored on a magnetic or optical medium. A large storage capacity is required, because it takes about 45,000 bytes to store a typical compressed text page of 2,500 characters and as much as 1,000,000 bytes to store a page containing an image. Aside from document imaging applications, digital scanning is used for transmission of documents via facsimile, in satellite photography, and in other applications.

An image scanner digitizes an entire document page for storage and display as an image and does not recognize characters and words of text. The stored material

therefore cannot be linguistically manipulated by text processing and other software techniques. When such manipulation is desired, a software program performs the optical character recognition (OCR) function by converting each optically scanned character into an electric signal and comparing it with the internally stored representation of an alphabet of characters, so as to select from it the one that matches the scanned character most closely or to reject it as an unidentifiable token. The more sophisticated of present-day OCR programs distinguish shapes, sizes, and pitch of symbols - including handwriting - and learn from experience. A universal optical character recognition machine is not available, however, for even a single alphabet.

Still photographs can be digitized by scanning or transferred from film to a compact digital disc holding more than 100 images. A recent development, the digital camera, makes it possible to bypass the film/paper step completely by capturing the image into the camera's random-access memory or a special diskette and then transferring it to a personal computer. Since both technologies produce a graphics file, in either case the image is editable by means of suitable software.

The digital recording of sound is important, because speech is the most frequently used natural carrier of communicable information. Direct capture of sound into personal computers is accomplished by means of a digital signal processor (DSP) chip, a special-purpose device built into the computer to perform array-processing operations. Conversion of analog audio signals to digital recordings is a commonplace process that has been used for years by the telecommunications and entertainment industries. Although the resulting digital sound track can be edited, automatic speech recognition—analogous to the recognition of characters and words in text by means of optical character recognition - is still under development. When perfected, voice recognition is certain to have a tremendous impact on the way humans communicate with recorded information, with computers, and among themselves.

By the beginning of the 1990s, the technology to record (or convert), store in digital form, and edit all visually and aurally perceived signals - text, graphics, still

images, animation, motion video, and sound - had thus become available and affordable. These capabilities have opened a way for a new kind of multimedia document that employs print, video, and sound to generate more powerful and colourful messages, communicate them securely at electronic speeds, and allow them to be modified almost at will. The traditional business letter, newspaper, journal, and book will no longer be the same.

INVENTORY OF RECORDED INFORMATION

The development of recording media and techniques enabled society to begin building a store of human knowledge. The idea of collecting and organizing written records is thought to have originated in Sumer about 5,000 years ago; Egyptian writing was introduced soon after. Early collections of Sumerian and Egyptian writings, recorded in cuneiform on clay tablets and in hieroglyphic script on papyrus, contained information about legal and economic transactions. In these and other early document collections (e.g., those of China produced during the Shang dynasty in the 2nd millennium BC and Buddhist collections in India dating to the 5th century BC), it is difficult to separate the concepts of the archive and the library.

From the Middle East the concept of document collections penetrated the Greco-Roman world. Roman kings institutionalized the population and property census as early as the 6th century BC. The great Library of Alexandria, established in the 3rd century BC, is best known as a large collection of papyri containing inventories of property, taxes, and other payments by citizens to their rulers and to each other. It is, in short, the ancient equivalent of today's administrative information systems.

The scholarly splendour of the Islamic world from the 8th to 13th century AD can in large part be attributed to the maintenance of public and private book libraries. The Bayt al-Hikmah ("House of Wisdom"), founded in AD 830 in Baghdad, contained a public library with a large collection of materials on a wide range of subjects, and the 10th-century library of Caliph al- Hakam in Cordova, Spain, boasted more than 400,000 books.

Primary and secondary literature.

The late but rapid development of European libraries from the 16th century on followed the invention of printing from movable type, which spurred the growth of the printing and publishing industries. Since the beginning of the 17th century, literature has become the principal medium for disseminating knowledge. The phrase primary literature is used to designate original information in various printed formats: newspapers, monographs, conference proceedings, learned and trade journals, reports, patents, bulletins, and newsletters. The scholarly journal, the classic medium of scientific communication, first appeared in 1665. Three hundred years later the number of periodical titles published in the world was estimated at more than 60,000, reflecting not only growth in the number of practitioners of science and expansion of its body of knowledge through specialization but also a maturing of the system of rewards that encourages scientists to publish.

The sheer quantity of printed information has for some time prevented any individual from fully absorbing even a minuscule fraction of it. Such devices as tables of contents, summaries, and indexes of various types, which aid in identifying and locating relevant information in primary literature, have been in use since the 16th century and led to the development of what is termed secondary literature during the 19th century. The purpose of secondary literature is to "filter" the primary information sources, usually by subject area, and provide the indicators to this literature in the form of reviews, abstracts, and indexes. Over the past 100 years there has evolved a system of disciplinary, national, and international abstracting and indexing services that acts as a gateway to several attributes of primary literature: authors, subjects, publishers, dates (and languages) of publication, and citations. The professional activity associated with these access-facilitating tools is called documentation.

The quantity of printed materials also makes it impossible, as well as undesirable, for any institution to acquire and house more than a small portion of it. The husbanding of recorded information has become a matter of public policy, as many countries have established national libraries and archives to direct the orderly

acquisition of analog-form documents and records. Since these institutions alone are not able to keep up with the output of such documents and records, new forms of cooperative planning and sharing recorded materials are evolving - namely, public and private, national and regional library networks and consortia.

Databases.

The emergence of digital technology in the mid-20th century has affected humankind's inventory of recorded information dramatically. During the early 1960s computers were used to digitize text for the first time; the purpose was to reduce the cost and time required to publish two American abstracting journals, the Index Medicus of the National Library of Medicine and the Scientific and Technical Aerospace Reports of the National Aeronautics and Space Administration (NASA). By the late 1960s such bodies of digitized alphanumeric information, known as bibliographic and numeric databases, constituted a new type of information resource. This resource is husbanded outside the traditional repositories of information (libraries and archives) by database "vendors." Advances in computer storage, telecommunications, software for computer sharing, and automated techniques of text indexing and searching fueled the development of an on-line database service industry. Meanwhile, electronic applications to bibliographic control in libraries and archives have led to the development of computerized catalogs and of union catalogs in library networks. They also have resulted in the introduction of comprehensive automation programs in these institutions.

The explosive growth of communications networks after 1990, particularly in the scholarly world, has accelerated the establishment of the "virtual library." At the leading edge of this development is public-domain information. Residing in thousands of databases distributed worldwide, a growing portion of this vast resource is now accessible almost instantaneously via the Internet, the web of computer networks linking the global communities of researchers and, increasingly, nonacademic organizations. Internet resources of electronic information include selected library catalogs, collected works of the literature,

some abstracting journals, full-text electronic journals, encyclopaedias, scientific data from numerous disciplines, software archives, demographic registers, daily news summaries, environmental reports, and prices in commodity markets, as well as hundreds of thousands of electronic-mail and bulletin-board messages.

The vast inventory of recorded information can be useful only if it is systematically organized and if mechanisms exist for locating in it items relevant to human needs. The main approaches for achieving such organization are reviewed in the following section, as are the tools used to retrieve desired information.

ORGANIZATION AND RETRIEVAL OF INFORMATION

In any collection, physical objects are related by order. The ordering may be random or according to some characteristic called a key. Such characteristics may be intrinsic properties of the objects (e.g., size, weight, shape, or colour), or they may be assigned from some agreed-upon set, such as object class or date of purchase. The values of the key are arranged in a sorting sequence that is dependent on the type of key involved: alphanumeric key values are usually sorted in alphabetic sequence, while other types may be sorted on the basis of similarity in class, such as books on a particular subject or flora of the same genus.

In most cases, order is imposed on a set of information objects for two reasons: to create their inventory and to facilitate locating specific objects in the set. There also exist other, secondary objectives for selecting a particular ordering, as, for example, conservation of space or economy of effort in fetching objects. Unless the objects in a collection are replicated, any ordering scheme is one-dimensional and unable to meet all the functions of ordering with equal effectiveness. The main approach for overcoming some of the limitations of one-dimensional ordering of recorded information relies on extended description of its content and, for analog-form information, of some features of the physical items. This approach employs various tools of content analysis that subsequently facilitate accessing and searching recorded information.

Description and content analysis of analog-form records.

The collections of libraries and archives, the primary repositories of analog-form information, constitute one-dimensional ordering of physical materials in print (documents), in image form (maps and photographs), or in audio-video format (recordings and videotapes). To break away from the confines of one-dimensional ordering, librarianship has developed an extensive set of attributes in terms of which it describes each item in the collection. The rules for assigning these attributes are called cataloging rules. Descriptive cataloging is the extraction of bibliographic elements (author names, title, publisher, date of publication, etc.) from each item; the assignment of subject categories or headings to such items is termed subject cataloging.

Conceptually, the library catalog is a table or matrix in which each row describes a discrete physical item and each column provides values of the assigned key. When such a catalog is represented digitally in a computer, any attribute can serve as the ordering key. By sorting the catalog on different keys, it is possible to produce a variety of indexes as well as subject bibliographies. More importantly, any of the attributes of a computerized catalog becomes a search key (access point) to the collection, surpassing the utility of the traditional card catalog.

The most useful access key to analog-form items is subject. The extensive lists of subject headings of library classification schemes provide, however, only a gross access tool to the content of the items. A technique called indexing provides a refinement over library subject headings. It consists of extracting from the item or assigning to it subject and other "descriptors" - words or phrases denoting significant concepts (topics, names) that occur in or characterize the content of the record. Indexing frequently accompanies abstracting, a technique for condensing the full text of a document into a short summary that contains its main ideas (but invariably incurs an information loss and often introduces a bias). Computer-printed, indexed abstracting journals provide a means of keeping users informed of primary information sources.

Description and content analysis of digital-form information.

The description of an electronic document generally follows the principles of bibliographic cataloging if the document is part of a database that is expected to be accessed directly and individually. When the database is an element of a universe of globally distributed database servers that are searchable in parallel, the matter of document naming is considerably more challenging, because several complexities are introduced. The document description must include the name of the database server - i.e., its physical location. Because database servers may delete particular documents, the description must also contain a pointer to the document's logical address (the generating organization). In contrast to their usefulness in the descriptive cataloging of analog documents, physical attributes such as format and size are highly variable in the milieu of electronic documents and therefore are meaningless in a universal document-naming scheme. On the other hand, the data type of the document (text, sound, etc.) is critical to its transmission and use. Perhaps the most challenging design is the "living document" - a constantly changing pastiche consisting of sections electronically copied from different documents, interspersed with original narrative or graphics or voice comments contributed by persons in distant locations, whose different versions reside on different servers. Efforts are under way to standardize the naming of documents in the universe of electronic networks.

Machine indexing.

The subject analysis of electronic text is accomplished by means of machine indexing, using one of two approaches: the assignment of subject descriptors from an unlimited vocabulary (free indexing) or their assignment from a list of authorized descriptors (controlled indexing). A collection of authorized descriptors is called an authority list or, if it also displays various relationships among descriptors such as hierarchy or synonymy, a thesaurus. The result of the indexing process is a computer file known as an inverted index, which is an alphabetic listing of descriptors and the addresses of their occurrence in the document body.

Full-text indexing, the use of every character string (word of a natural language) in the text as an index term, is an extreme case of free-text indexing: each word in the

document (except function words such as articles and prepositions) becomes an access point to it. Used earlier for the generation of concordances in literary analysis and other computer applications in the humanities, full-text indexing placed great demands on computer storage because the resulting index is at least as large as the body of the text. With decreasing cost of mass storage, automatic full-text indexing capability has been incorporated routinely into state-of-the-art information-management software.

Text indexing may be supplemented by other syntactic techniques, so as to increase its precision or robustness. One such method, the Standard Generalized Markup Language (SGML), takes advantage of standard text markers used by editors to pinpoint the location and other characteristics of document elements (paragraphs and tables, for example). In indexing spatial data such as maps and astronomical images, the textual index specifies the search areas, each of which is further described by a set of coordinates defining a rectangle or irregular polygon. These digital spatial document attributes are then used to retrieve and display a specific point or a selected region of the document. There are other specialized techniques that may be employed to augment the indexing of specific document types, such as encyclopaedias, electronic mail, catalogs, bulletin boards, tables, and maps.

Semantic content analysis.

The analysis of digitally recorded natural-language information from the semantic viewpoint is a matter of considerable complexity, and it lies at the foundation of such incipient applications as automatic question answering from a database or retrieval by means of unrestricted natural-language queries. The general approach has been that of computational linguistics: to derive representations of the syntactic and semantic relations among the linguistic elements of sentences and larger parts of the document. Syntactic relations are described by parsing (decomposing) the grammar of sentences. For semantic representation, three related formalisms dominate. In a so-called semantic network, conceptual entities such as objects, actions, or events are represented as a graph of linked nodes (Figure 5). "Frames"

represent, in a similar graph network, physical or abstract attributes of objects and in a sense define the objects. In "scripts," events and actions rather than objects are defined in terms of their attributes.

Indexing and linguistic analyses of text generate a relatively gross measure of the semantic relationship, or subject similarity, of documents in a given collection. Subject similarity is, however, a pragmatic phenomenon that varies with the observer and the circumstances of an observation (purpose, time, and so forth). A technique experimented with briefly in the mid-1960s, which assigned to each document one or more "roles" (functions) and one or more "links" (pointers to other documents having the same or a similar role), showed potential for a pragmatic measure of similarity; its use, however, was too unwieldy for the computing environment of the day. Some 20 years later, a similar technique became popular under the name "hypertext." In this technique, documents that a person or a group of persons consider related (by concept, sequence, hierarchy, experience, motive, or other characteristics) are connected via "hyperlinks," mimicking the way humans associate ideas. Objects so linked need not be only text; speech and music, graphics and images, and animation and video can all be interlinked into a "hypermedia" database. The objects are stored with their hyperlinks, and a user can easily navigate the network of associations by clicking with a mouse on a series of entries on a computer screen. Another technique that elicits semantic relationships from a body of text is SGML.

Image analysis.

The content analysis of images is accomplished by two primary methods: image processing and pattern recognition. Image processing is a set of computational techniques for analyzing, enhancing, compressing, and reconstructing images. Pattern recognition is an information-reduction process: the assignment of visual or logical patterns to classes based on the features of these patterns and their relationships. The stages in pattern recognition involve measurement of the object to identify distinguishing attributes, extraction of features for the defining attributes, and assignment of the object to a class based on these features. Both

image processing and pattern recognition have extensive applications in various areas, including astronomy, medicine, industrial robotics, and remote sensing by satellites.

Speech analysis.

The immediate objective of content analysis of digital speech is the conversion of discrete sound elements into their alphanumeric equivalents. Once so represented, speech can be subjected to the same techniques of content analysis as natural-language text - i.e., indexing and linguistic analysis. Converting speech elements into their alphanumeric counterparts is an intriguing problem because the "shape" of speech sounds embodies a wide range of many acoustic characteristics and because the linguistic elements of speech are not clearly distinguishable from one another. The technique used in speech processing is to classify the spectral representations of sound and to match the resulting digital spectrographs against prestored "templates" so as to identify the alphanumeric equivalent of the sound. (The obverse of this technique, the digital-to-analog conversion of such templates into sound, is a relatively straightforward approach to generating synthetic speech.) Speech processing is complex as well as expensive in terms of storage capacity and computational requirements. State-of-the-art speech recognition systems can identify limited vocabularies and parts of distinctly spoken speech and can be programmed to recognize tonal idiosyncracies of individual speakers. When more robust and reliable techniques become available and the process is made computationally tractable (as is expected with parallel computers), humans will be able to interact with computers via spoken commands and queries on a routine basis. In many situations this may make the keyboard obsolete as a data-entry device.

Storage structures for digital-form information.

Digital information is stored in complex patterns that make it feasible to address and operate on even the smallest element of symbolic expression, as well as on larger strings such as words or sentences and on images and sound.

From the viewpoint of digital information storage, it is useful to distinguish between "structured" data, such as inventories of objects that can be represented by short symbol strings and numbers, and "unstructured" data, such as the natural-language text of documents or pictorial images. The principal objective of all storage structures is to facilitate the processing of data elements based on their relationships; the structures thus vary with the type of relationship they represent. The choice of a particular storage structure is governed by the relevance of the relationships it allows to be represented to the information-processing requirements of the task or system at hand.

In information systems whose store consists of unstructured databases of natural-language records, the objective is to retrieve records (or portions thereof) based on the presence in the records of words or short phrases that constitute the query. Since there exists an index as a separate file that provides information about the locations of words and phrases in the database records, the relationships that are of interest (e.g., word adjacency) can be calculated from the index. Consequently, the database text itself can be stored as a simple ordered sequential file of records. The majority of the computations use the index, and they access the text file only to pull out the records or those portions that satisfy the result of the computations. The sequential file structure remains popular, with document-retrieval software intended for use with personal computers and CD-ROM databases.

When relationships among data elements need to be represented as part of the records so as to make more efficient the desired operations on these records, two types of "chained" structures are commonly used: hierarchical and network. In the hierarchical file structure, records are arranged in a scheme resembling a family tree, with records related to one another from top to bottom. In the network file structure, records are arranged in groupings known as sets; these can be connected in any number of ways, giving rise to considerable flexibility. In both hierarchical and network structures, the relationships are shown by means of "pointers" (i.e., identifiers such as addresses or keys) that become part of the records.

Another type of database storage structure, the relational structure, has become increasingly popular since the late 1970s. Its major advantage over the hierarchical and network structures is the ability to handle unanticipated data relationships without pointers. Relational storage structures are two-dimensional tables consisting of rows and columns, much like the conceptual library catalog mentioned above. The elegance of the relational model lies in its conceptual simplicity, the availability of theoretical underpinnings (relational algebra), and the ability of its associated software to handle data relationships without the use of pointers. The relational model was initially used for databases containing highly structured information. In the 1990s it has largely replaced the hierarchical and network models, and it has also become the model of choice for large-scale information-management applications, both textual and multimedia.

The feasibility of storing large volumes of full text on an economic medium (the digital optical disc) has renewed interest in the study of storage structures that permit more powerful retrieval and processing techniques to operate on cognitive entities other than words, to facilitate more extensive semantic content and context analysis, and to organize text conceptually into logical units rather than those dictated by printing conventions.

Query languages.

The uses of databases are manifold. They provide a means of retrieving records or parts of records and performing various calculations before displaying the results. The interface by which such manipulations are specified is called the query language. Whereas early query languages were originally so complex that interacting with electronic databases could be done only by specially trained individuals, recent interfaces are more user-friendly, allowing casual users to access database information.

The main types of popular query modes are the "menu," the"fill-in-the-blank" technique, and the structured query. Particularly suited for novices, the menu requires a person to choose from several alternatives displayed on the video terminal screen. The fill-in-the-blank technique is one in which the user is

prompted to enter key words as search statements. The structured query approach is effective with relational databases. It has a formal, powerful syntax that is in fact a programming language, and it is able to accommodate logical operators. One implementation of this approach, the Structured Query Language (SQL), has the form

select [field Fa, Fb, . . . , Fn]
from [database Da, Db, . . . , Dn]
where [field Fa = abc] and [field Fb = def].

Structured query languages support database searching and other operations by using commands such as "find," "delete," "print," "sum," and so forth. The sentencelike structure of an SQL query resembles natural language except that its syntax is limited and fixed. Instead of using an SQL statement, it is possible to represent queries in tabular form. The technique, referred to as query-by-example (or QBE), displays an empty tabular form and expects the searcher to enter the search specifications into appropriate columns. The program then constructs an SQL-type query from the table and executes it.

The most flexible query language is of course natural language. The use of natural-language sentences in a constrained form to search databases is allowed by some commercial database management software. These programs parse the syntax of the query; recognize its action words and their synonyms; identify the names of files, records, and fields; and perform the logical operations required. Experimental systems that accept such natural-language queries in spoken voice have been developed; however, the ability to employ unrestricted natural language to query unstructured information will require further advances in machine understanding of natural language, particularly in techniques of representing the semantic and pragmatic context of ideas. The prospect of an intelligent conversation between humans and a large store of digitally encoded knowledge is not imminent.

Information searching and retrieval.

State-of-the-art approaches to retrieving information employ two generic techniques: (1) matching words in the query against the database index (key-word

searching) and (2) traversing the database with the aid of hypertext or hypermedia links.

Key-word searches can be made either more general or more narrow in scope by means of logical operators (e.g., disjunction and conjunction). Because of the semantic ambiguities involved in free-text indexing, however, the precision of the key-word retrieval technique - that is, the percentage of relevant documents correctly retrieved from a collection - is far from ideal, and various modifications have been introduced to improve it. In one such enhancement, the search output is sorted by degree of relevance, based on a statistical match between the key words in the query and in the document; in another, the program automatically generates a new query using one or more documents considered relevant by the user. Key-word searching has been the dominant approach to text retrieval since the early 1960s; hypertext has so far been largely confined to personal or corporate information-retrieval applications.

The exponential growth of the use of computer networks in the 1990s presages significant changes in systems and techniques of information retrieval. In a wide-area information service, a number of which began operating at the beginning of the 1990s on the Internet computer network, a user's personal computer or terminal (called a client) can search simultaneously a number of databases maintained on heterogeneous computers (called servers). The latter are located at different geographic sites, and their databases contain different data types and often use incompatible data formats. The simultaneous, distributed search is possible because clients and servers agree on a standard document addressing scheme and adopt a common communications protocol that accommodates all the data types and formats used by the servers. Communication with other wide-area services using different protocols is accomplished by routing through so-called gateways capable of protocol translation. Several representative clients are shown: a "dumb" terminal (i.e., one with no internal processor), a personal computer (PC), and Macintosh (trademark; Mac), and NeXT (trademark) machines. They have access to data on the servers sharing a common protocol as well as to data provided by

services that require protocol conversion via the gateways. Network news is such a wide-area service, containing hundreds of news groups on a variety of subjects, by which users can read and post messages.

Evolving information-retrieval techniques, exemplified by an experimental interface to the NASA space shuttle reference manual, combine natural language, hyperlinks, and key-word searching. Other techniques, seeking higher levels of retrieval precision and effectiveness, are studied by researchers involved with artificial intelligence and neural networks. The next major milestone may be a computer program that traverses the seamless information universe of wide-area electronic networks and continuously filters its contents through profiles of organizational and personal interest: the information robot of the 21st century.

INFORMATION DISPLAY

For humans to perceive and understand information, it must be presented as print and image on paper; as print and image on film or on a video terminal; as sound via radio or telephony; as print, sound, and video in motion pictures, on television broadcasts, or at lectures and conferences; or in face-to-face encounters. Except for live encounters and audio information, such displays emanate increasingly from digitally stored data, with the output media being video, print, and sound.

Video.

Possibly the most widely used video display device, at least in the industrialized world, is the television set. Designed primarily for video and sound, its image resolution is inadequate for alphanumeric data except in relatively small amounts. Use of the television set in text-oriented information systems has been limited to menu-oriented applications such as videotex, in which information is selected from hierarchically arranged menus (with the aid of a numeric keyboard attachment) and displayed in fixed frames. The television, computer, and communications technologies are, however, converging in a high-resolution digital television set capable of receiving alphanumeric, video, and audio signals.

The computer video terminal is today's ubiquitous interface that transforms computer-stored data into analog form for human viewing. The two basic

apparatuses used are the cathode-ray tube (CRT) and the more recent flat-panel display. In CRT displays an electron gun emits beams of electrons on a phosphorus-coated surface; the beams are deflected, forming visible patterns representative of data. Flat-panel displays use one of four different media for visual representation of data: liquid crystal, light-emitting diodes, plasma panels, and electroluminescence. Advanced video display systems enable the user to scroll, page, zoom (change the scale of the details of the display image for enhancement), divide the screen into multiple colours and windows (viewing areas), and in some cases even activate commands by touching the screen instead of using the keyboard. The information capacity of the terminal screen depends on its resolution, which ranges from low (character-addressable) to high (bit-addressable). High resolution is indispensable for the display of graphic and video data in state-of-the-art workstations, such as those used in engineering or information systems design.

Print.

Modern society continues to be dominated by printed information. The convenience and portability of print on paper make it difficult to imagine the paperless world that some have predicted. The generation of paper print has changed considerably, however. Although manual typesetting is still practiced for artwork, in special situations, and in some developing countries, electronic means of composing pages for subsequent reproduction by photoduplication and other methods has become commonplace.

Since the 1960s, volume publishing has become an automated process using large computers and high-speed printers to transfer digitally stored data on paper. The appearance of microcomputer-based publishing systems has proved to be another significant advance. Economical enough to allow even small organizations to become in-house publishers, these so-called desktop publishing systems are able to format text and graphics interactively on a high-resolution video screen with the aid of page-description command languages. Once a page has been formatted, the entire image is transferred to an electronic printing or photocomposition device.

Printers.

Computer printers are commonly divided into two general classes according to the way they produce images on paper: impact and nonimpact. In the first type, images are formed by the print mechanism making contact with the paper through an ink-coated ribbon. The mechanism consists either of print hammers shaped like characters or of a print head containing a row of pins that produce a pattern of dots in the form of characters or other images.

Most nonimpact printers form images from a matrix of dots, but they employ different techniques for transferring images to paper. The most popular type, the laser printer, uses a beam of laser light and a system of optical components to etch images on a photoconductor drum from which they are carried via electrostatic photocopying to paper. Light-emitting diode (LED) printers resemble laser printers in operation but direct light from energized diodes rather than a laser onto a photoconductive surface. Ion-deposition printers make use of technology similar to that of photocopiers for producing electrostatic images. Another type of nonimpact printer, the ink-jet printer, sprays electrically charged drops of ink onto the print surface.

Microfilm and microfiche.

Alphanumeric and image information can be transferred from digital computer storage directly to film. Reel microfilm and microfiche (a flat sheet of film containing multiple microimages reduced from the original) were popular methods of document storage and reproduction for several decades. During the 1990s they have been largely replaced by optical disc technology (see above Recording media).

Voice.

In synthetic speech generation, digitally prestored sound elements are converted to analog sound signals and combined to form words and sentences. Digital-to-analog converters are available as inexpensive boards for microcomputers or as software for larger machines. Human speech is the most effective natural form of communication, and so applications of this technology are becoming increasingly

popular in situations where there are numerous requests for specific information (e.g., time, travel, and entertainment), where there is a need for repetitive instruction, in electronic voice mail (the counterpart of electronic text mail), and in toys.

DISSEMINATION OF INFORMATION

The process of recording information by handwriting was obviously laborious and required the dedication of the likes of Egyptian scribes or monks in monasteries around the world. It was only after mechanical means of reproducing writing were invented that information records could be duplicated more efficiently and economically.

The first practical method of reproducing writing mechanically was block printing; it was developed in China during the T'ang dynasty (618-907). Ideographic text and illustrations were engraved in wooden blocks, inked, and copied on paper. Used to produce books as well as cards, charms, and calendars, block printing spread to Korea and Japan but apparently not to the Islamic or European Christian civilizations. European woodcuts and metal engravings date only to the 14th century.

Printing from movable type was also invented in China (in the mid-11th century AD). There and in the bookmaking industry of Korea, where the method was applied more extensively during the 15th century, the ideographic type was made initially of baked clay and wood and later of metal. The large number of typefaces required for pictographic text composition continued to handicap printing in the Orient until the present time.

The invention of character-oriented printing from movable type (1440-50) is attributed to the German printer Johannes Gutenberg. Within 30 years of his invention, the movable-type printing press was in use throughout Europe. Character-type pieces were metallic and apparently cast from metallic molds; paper and vellum (calfskin parchment) were used to carry the impressions. Gutenberg's technique of assembling individual letters by hand was employed until 1886, when the German-born American printer Ottmar Mergenthaler developed

the Linotype, a keyboard-driven device that cast lines of type automatically. Typesetting speed was further enhanced by the Monotype technique, in which a perforated paper ribbon, punched from a keyboard, was used to operate a type-casting machine.

Mechanical methods of typesetting prevailed until the 1960s. Since that time they have been largely supplanted by the electronic and optical printing techniques described in the previous section.

Unlike the use of movable type for printing text, early graphics were reproduced from wood relief engravings in which the nonprinting portions of the image were cut away. Musical scores, on the other hand, were reproduced from etched stone plates. At the end of the 18th century the German printer Aloys Senefelder developed lithography, a planographic technique of transferring images from a specially prepared surface of stone. In offset lithography the image is transferred from zinc or aluminum plates instead of stone, and in photoengraving such plates are superimposed with film and then etched.

The first successful photographic process, the daguerreotype, was developed during the 1830s. The invention of photography, aside from providing a new medium for capturing still images and later video in analog form, was significant for two other reasons. First, recorded information (textual and graphic) could be easily reproduced from film and, second, the image could be enlarged or reduced. Document reproduction from film to film has been relatively unimportant, because both printing and photocopying (see above) are cheaper. The ability to reduce images, however, has led to the development of the microform, the most economical method of disseminating analog-form information.

Another technique of considerable commercial importance for the duplication of paper-based information is photocopying, or dry photography. Printing is most economical when large numbers of copies are required, but photocopying provides a fast and efficient means of duplicating records in small quantities for personal or local use. Of the several technologies that are in use, the most popular process, xerography, is based on electrostatics.

While the volume of information issued in the form of printed matter continues unabated, the electronic publishing industry has begun to disseminate information in digital form. The digital optical disc (see above Recording media) is developing as an increasingly popular means of issuing large bodies of archival information - for example, legislation, court and hospital records, encyclopaedias and other reference works, referral databases, and libraries of computer software. Full-text databases, each containing digital page images of the complete text of some 400 periodicals stored on CD-ROM, entered the market in 1990. The optical disc provides the mass production technology for publication in machine-readable form. It offers the prospect of having large libraries of information available in virtually every school and at many professional workstations.

The coupling of computers and digital telecommunications is also changing the modes of information dissemination. High-speed digital satellite communications facilitate electronic printing at remote sites; for example, the world's major newspapers and magazines transmit electronic page copies to different geographic locations for local printing and distribution. Updates of catalogs, computer software, and archival databases are distributed via electronic mail, a method of rapidly forwarding and storing bodies of digital information between remote computers.

Indeed, a large-scale transformation is taking place in modes of formal as well as informal communication. For more than three centuries, formal communication in the scientific community has relied on the scholarly and professional periodical, widely distributed to tens of thousands of libraries and to tens of millions of individual subscribers. In 1992 a major international publisher announced that its journals would gradually be available for computer storage in digital form; and in that same year the State University of New York at Buffalo began building a completely electronic, paperless library. The scholarly article, rather than the journal, is likely to become the basic unit of formal communication in scientific disciplines; digital copies of such an article will be transmitted electronically to subscribers or, more likely, on demand to individuals and organizations who learn

of its existence through referral databases and new types of alerting information services. The Internet already offers instantaneous public access to vast resources of noncommercial information stored in computers around the world.

Similarly, the traditional modes of informal communications - various types of face-to-face encounters such as meetings, conferences, seminars, workshops, and classroom lectures - are being supplemented and in some cases replaced by electronic mail, electronic bulletin boards (a technique of broadcasting newsworthy textual and multimedia messages between computer users), and electronic teleconferencing and distributed problem-solving (a method of linking remote persons in real time by voice-and-image communication and special software called "groupware"). These technologies are forging virtual societal networks - communities of geographically dispersed individuals who have common professional or social interests.

Information systems

The primary vehicles for the purposeful, orchestrated processing of information are information systems - constructs that collect, organize, store, process, and display information in all its forms (raw data, interpreted data, knowledge, and expertise) and formats (text, video, and voice). In principle, any record-keeping system - e.g., an address book or a train schedule - may be regarded as an information system. What sets modern information systems apart is their electronic dimension, which permits extremely fast, automated manipulation of digitally stored data and their transformation from and to analog representation.

IMPACT OF INFORMATION TECHNOLOGY

Electronic information systems are a phenomenon of the second half of the 20th century. Their evolution is closely tied with advances in two basic technologies: integrated circuits and digital communications.

Integrated circuits are silicon chips containing transistors that store and process information. Advances in the design of these chips, which were first developed in 1958, are responsible for an exponential increase in the cost performance of computer components. For more than two decades the capacity of the basic

integrated circuit, the dynamic random-access memory (DRAM) chip, has doubled consistently in intervals of less than two years: from 1,000 transistors (1 kilobit) per chip in 1970 to 1,000,000 (1 megabit) in 1987, 16 megabits in 1993, and 1,000,000,000 (1 gigabit) predicted for the year 2000. A gigabit chip has the capacity of 125,000,000 bytes, approximately equivalent to 14,500 pages, or more than 12 volumes, of Encyclopædia Britannica.

The speed of microprocessor chips, measured in millions of instructions per second (MIPS), is also increasing near-exponentially: from 10 MIPS in 1985 to 100 MIPS in 1993, with 1,000 MIPS predicted for 1995. By the year 2000 a single chip may process 64 billion instructions per second. If in a particular computing environment in 1993 a chip supported 10 simultaneous users, in the year 2000 such a chip could theoretically support several thousand users.

Full exploitation of these developments for the realm of information systems requires comparable advances in software disciplines. Their major contribution has been to open the use of computer technology to persons other than computer professionals. Interactive applications in the office and home have been made possible by the development of easy-to-use software products for the creation, maintenance, manipulation, and querying of files and records. The database has become a central organizing framework for many information systems, taking advantage of the concept of data independence, which allows data sharing among diverse applications. Database management system (DBMS) software today incorporates high-level programming facilities that do not require one to specify in detail how the data should be processed. The programming discipline as a whole, however, progresses in an evolutionary manner. Whereas semiconductor field advances are measured by orders of magnitude, the writing and understanding of large suites of software that characterize complex information systems progress more slowly. The complexity of the data processes that comprise very large information systems has so far eluded major breakthroughs, and the cost-effectiveness of the software development sector improves only gradually.

The utility of computers is vastly augmented by their ability to communicate with one another, so as to share data and its processing. Local-area networks (LANs) permit the sharing of data, programs, printers, and electronic mail within offices and buildings. In wide-area networks, such as the Internet, which connect thousands of computers around the globe, computer-to-computer communication uses a variety of media as transmission lines - electric-wire audio circuits, coaxial cables, radio and microwaves (as in satellite communication), and, most recently, optical fibres. The latter are replacing coaxial cable in the Integrated Services Digital Network (ISDN), which is capable of carrying digital information in the form of voice, text, and video simultaneously. To communicate with another machine, a computer requires data circuit-terminating equipment, or DCE, which connects it to the transmission line. When an analog line such as a dial-up telephone line is used, the DCE is called a modem (for modulator/demodulator); it also provides the translation of the digital signal to analog and vice versa. By using data compression, the relatively inexpensive high-speed modems currently in use can transmit data at speeds of more than 100 kilobits per second. When digital lines are used, the DCE allows substantially higher speeds; for instance, the U.S. scholarly network NSFNET, set up by the National Science Foundation, transmits information at 45 million bits per second. The National Research and Education Network, proposed by the U.S. government in 1991, is designed to send data at speeds in the gigabit-per-second range, comfortably moving gigantic volumes of text, video, and sound across a web of digital highways.

Computer networks are complex entities. Each network operates according to a set of procedures called the network protocol. The proliferation of incompatible protocols during the early 1990s has been brought under relative control by the Open Systems Interconnection (OSI) reference Model formulated by the International Organization for Standardization. To the extent that individual protocols conform to the OSI recommendations, computer networks can now be interconnected efficiently through gateways.

Computer networking facilitates the current trend toward distributed information systems. At the corporate level, the central database may be distributed over a number of computer systems in different locations, yet its querying and updating are carried out simultaneously against the composite database. An individual searching for public-access information can traverse disparate computer networks to peruse hundreds of autonomous databases and within seconds or minutes download a copy of the desired document into a personal workstation.

The future of information systems may be gleaned from several areas of current research. As all information carriers (text, video, and sound) can be converted to digital form and manipulated by increasingly sophisticated techniques, the ranges of media, functions, and capabilities of information systems are constantly expanding. Evolving techniques of natural-language processing and understanding, knowledge representation, and neural process modeling have begun to join the more traditional repertoire of methods of content analysis and manipulation. The use of these techniques opens the possibility of eliciting new knowledge from existing data, such as the discovery of a previously unknown medical syndrome or of a causal relationship in a disease. Computer visualization, a new field that has grown expansively since the early 1990s, deals with the conversion of masses of data emanating from instruments, databases, or computer simulations into visual displays - the most efficient method of human information reception, analysis, and exchange. Related to computer visualization is the research area of virtual reality or virtual worlds, which denotes the generation of synthetic environments through the use of three-dimensional displays and interaction devices. A number of research directions in this area are particularly relevant to future information systems: knowledge-based world modeling; the development of physical analogues for abstract quantitative and organizational data; and search and retrieval in large virtual worlds. The cumulative effect of these new research areas is a gradual transformation of the role of information systems from that of data processing to that of cognition aiding.

Present-day computers are remarkably versatile machines capable of assisting humans in nearly every problem-solving task that involves symbol manipulations. Television, on the other hand, has penetrated societies throughout the world as a noninteractive display device for combined video and audio signals. The impending convergence of three digital technologies - namely, the computer, very-high-definition television (V-HDTV), and ISDN data communications - is all but inevitable. In such a system, a large-screen multimedia display monitor, containing a 64-megabit primary memory and a billion-byte hard disk for data storage and playback, would serve as a computer and, over ISDN fibre links, an interactive television receiver.

ANALYSIS AND DESIGN OF INFORMATION SYSTEMS

The building of information systems falls within the domain of engineering. As is true with other engineering disciplines, the nature and tools of information systems engineering are evolving owing to both technological developments and better perceptions of societal needs for information services. Early information systems were designed to be operated by information professionals, and they frequently did not attain their stated social purpose. Modern information systems are increasingly used by persons who have little or no previous hands-on experience with information technology but who possess a much better perception about what this technology should accomplish in their professional and personal environments. A correct understanding of the requirements, preferences, and "information styles" of these end users is crucial to the design and success of today's information systems.

The methodology involved in building an information system consists of a set of iterative activities that are cumulatively referred to as the system's life cycle. The principal objective of the systems analysis phase is the specification of what the system is required to do. In the systems design phase such specifications are converted to a hierarchy of increasingly detailed charts that define the data required and decompose the processes to be carried out on data to a level at which they can be expressed as instructions of a computer program. The systems development phase consists of writing and testing computer software and of

developing data input and output forms and conventions. Systems implementation is the installation of a physical system and the activities it entails, such as the training of operators and users. Systems maintenance refers to the further evolution of the functions and structure of a system that results from changing requirements and technologies, experience with the system's use, and fine-tuning of its performance.

Many information systems are implemented with generic, "off-the-shelf" software rather than with custom-built programs; versatile database management software and its nonprocedural programming languages fit the needs of small and large systems alike. The development of large systems that cannot use off-the-shelf software is an expensive, time-consuming, and complex undertaking. Prototyping, an interactive session in which users confirm a system's proposed functions and features early in the design stage, is a practice intended to raise the probability of success of such an undertaking. Some of the tools of computer-aided software engineering available to the systems analyst and designer verify the logic of systems design, automatically generate a program code from low-level specifications, and automatically produce software and system specifications. The eventual goal of information systems engineering is to develop software "factories" that use natural language and artificial intelligence techniques as part of an integrated set of tools to support the analysis and design of large information systems.

CATEGORIES OF INFORMATION SYSTEMS

A taxonomy of information systems is not easily developed, because of their diversity and continuing evolution in structure and function. Earlier distinctions - manual versus automated, interactive versus off-line, real-time versus batch-processing - are no longer appropriate. A more frequently made distinction is in terms of application: use in business offices, factories, hospitals, and so on. In the functional approach taken in this article, information systems may be divided into two categories: organizational systems and public information utilities. Information systems in formal organizations may be further distinguished according to their

main purpose: support of managerial and administrative functions or support of operations and services. The former serve internal functions of the organizations, while the latter support the purposes for which these organizations exist.

Management-oriented information systems.

The most important functions that top executives perform include setting policies, planning, and preparing budgets. At the strategic level, these decision-making functions are supported by executive information systems. The objective of these systems is to gather, analyze, and integrate internal (corporate) and external (public) data into dynamic profiles of key corporate indicators. Depending on the nature of the organization's business, such indicators may relate to the status of high-priority programs, health of the economy, inventory and cash levels, performance of financial markets, relevant efforts of competitors, utilization of manpower, legislative events, and so forth. The indicators are displayed as text, tables, graphics, or time series, and optional access is provided to more detailed data. The data emanate not only from within the organization's production and administrative departments but also from external information sources, such as public databases (Figure 8). Present-day efforts, drawing on research in neural computers and networks, are to enhance executive information systems with adaptive and self-organizing abilities by means of learning from the executives' changing information needs and uses.

In military organizations, the approximate equivalent of executive information systems is command-and-control systems. Their purpose is to maintain control over some domain and, if needed, initiate corrective action. Their key characteristic is the real-time nature of the monitoring and decision-making functions. A command-and-control system typically assumes that the environment exercises pressure on the domain of interest (say, a naval force); the system then monitors the environment (collects intelligence data), analyzes the data, compares it with the desired state of the domain, and suggests actions to be taken. Systems of this kind are used at both strategic and tactical levels.

Both executive and military command-and-control systems make use of computational aids for data classification, modeling, and simulation. These capabilities are characteristic of a decision-support system (DSS), a composite of computer techniques for supporting executive decision making in relatively unstructured problem situations. Decision-support software falls into one of two categories: decision-aid programs, in which the decision maker assigns weighted values to every factor in the decision, and decision-modeling programs, in which the user explores different strategies to arrive at the desired outcome.

Administration-oriented information systems.

Administrative functions in formal organizations have as their objective the husbanding and optimization of corporate resources - namely, employees and their activities, inventories of materials and equipment, facilities, and finances. Administrative information systems support this objective. Commonly called management information systems (MIS), they focus primarily on resource administration and provide top management with reports of aggregate data. Executive information systems may be viewed as an evolution of administrative information systems in the direction of strategic tracking, modeling, and decision making.

Typically, administrative information systems consist of a number of modules, each supporting a particular function. The modules share a common database whose contents may, however, be distributed over a number of machines and locations. Financial information systems have evolved from the initial applications of punched cards before World War II to integrated accounting and finance systems that cover general accounting, accounts receivable and payable, payroll, purchasing, inventory control, and financial statements such as balance sheets. Functionally close to payroll systems are personnel information systems, which support the administration of the organization's human resources. Job and salary histories, inventory of skills, performance reviews, and other types of personnel data are combined in the database to assist personnel administration, explore potential effects of reorganization or new salary scales (or changes in benefits), and

match job requirements with skills. Project management information systems concentrate on resource allocation and task completion of organized activities; they usually incorporate such scheduling methods as the critical path method (CPM) or program evaluation and review technique (PERT).

Since the advent of microcomputers, information processing in organizations has become heavily supported by office automation tools. These involve six basic applications: text processing, database, spreadsheet, graphics, communications, and networking. Administrative systems in smaller organizations are usually built as extensions of office automation tools; in large organizations these tools form an interface to custom software. The current trend in office automation is toward integrating the first five applications into a software utility, either delivered to each microprocessor workstation from a "server" on the corporate computer network or integrated into other applications software.

Administrative information systems abound in organizations in both the private and public sectors throughout the industrialized world. In the retail industry, point-of-sale terminals are linked into distributed administrative information systems that contain financial and inventory modules at the department, store, geographic area, and corporate chain levels, with modeling facilities that help to determine marketing strategies and optimize profits. Administrative information systems are indispensable to government; the agencies of virtually all U.S. municipalities with more than 10,000 inhabitants use such systems. The systems are generally centred around a generic database management system and are increasingly supported by software modules and programs that permit data modeling - i.e., they acquire management orientation.

Service-oriented information systems.

Such information systems provide support for the operations or services that organizations perform for society. The systems are vertically oriented to specific sectors and industries (e.g., manufacturing, financial services, publishing, education, health, and entertainment). Rather than addressing management and administrative functions, they support activities and processes that are the reason

for an organization's existence - in most cases, some kind of manufacturing activity or the rendering of services. Systems of this kind vary greatly, but they tend to fall into three main types: manufacturing, transaction, and expert systems.

Computer-integrated manufacturing.

The conceptual goal of modern factories is computer-integrated manufacturing (CIM). The phrase denotes data-driven automation that affects all components of the manufacturing enterprise: design and development engineering, manufacturing, marketing and sales, and field support and service. Computer-aided design (CAD) systems were first applied in the electronics industry. Today they feature three-dimensional modeling techniques for drafting and manipulating solid objects on the screen and for deriving specifications for programs to drive numerical-control machines. Once a product is designed, its production process can be outlined using computer-aided process planning (CAPP) systems that help to select sequences of operations and machining conditions. Models of the manufacturing system can be simulated by computers before they are built. The basic manufacturing functions - machining, forming, joining, assembly, and inspection - are supported by computer-aided manufacturing (CAM) systems and automated materials-handling systems. Inventory control systems seek to maintain an optimal stock of parts and materials by tracking inventory movement, forecasting requirements, and initiating procurement orders.

Transaction-processing systems.

In nonmanufacturing service organizations the prevalent type of information system supports transaction processing. Transactions are sets of discrete inputs, submitted by users at unpredictable intervals, which call for database searching, analysis, and modification. The processor evaluates the request and executes it immediately. Portions of the processing function may be carried out at the intelligent terminal that originated the request so as to distribute the computational load. Response time (the elapsed time between the end of a request and the beginning of the reply) is an important characteristic of this type of real-time

teleprocessing system. Large transaction-processing systems often incorporate private telecommunications networks.

Teleprocessing transaction systems constitute the foundation of service industries such as banking, insurance, securities, transportation, and libraries. They are replacing the trading floor of the world's major stock exchanges, linking the latter via on-line telecommunications into a global financial market. Again, the core of a transaction system is its integrated database. The focus of the system is the recipient of services rather than the system operator. Because of this, a local travel agent is able to plan the complete itinerary of a traveler - including reservations for airlines, hotels, rental cars, cultural and sports performances, and even restaurants, on any continent - and to tailor these to the traveler's schedule and budget.

Expert systems.

A relatively new category of service-oriented information systems is the expert system, so called because its database stores a description of decision-making skills of human experts in some narrow domain of performance, such as medical image interpretation, taxation, brickwork design, configuration of computer system hardware, troubleshooting malfunctioning equipment, or beer brewing. The motivation for constructing expert systems is the desire to replicate the scarce, unstructured, and perhaps poorly documented empirical knowledge of specialists so that it can be readily used by others.

Expert systems have three components: (1) a software interface through which the user formulates queries by which the expert system solicits further information from the user and by which it explains to the user the reasoning process employed to arrive at an answer, (2) a database (called the knowledge base) consisting of axioms (facts) and rules for making inferences from these facts, and (3) a computer program (dubbed the inference engine) that executes the inference-making process. The knowledge base is a linked structure of rules that the human expert applies, often intuitively, in problem solving. The process of acquiring such knowledge typically has three phases: a functional analysis of the environment, users, and tasks performed by the expert; identification of concepts of the domain of expertise

and their classification according to various relationships; and an interview, by either human or automated techniques, of the expert (or experts) in action. The results of these steps are translated into so-called production rules (of the form "IF condition x exists, THEN action y follows") and stored in the knowledge base. Chains of production rules form the basis for the automated deductive capabilities of expert systems and for their ability to explain their actions to users.

Expert systems are a commercial variety of a class of computer programs called knowledge-based systems. Knowledge in expert systems is highly unstructured (i.e., the problem-solving process of the domain is not manifest), and it is stated explicitly in relationships or deductively inferred from the chaining of propositions. Since every condition that may be encountered must be described by a rule, rule-based expert systems cannot handle unanticipated events (but can evolve with usage) and remain limited to narrow problem domains.

Another variant of expert systems, one that does not possess this limitation, employs a knowledge base that consists of structured descriptions of real-world problem situations and of decisions actually made by human experts. In medicine, for example, the patient record contains descriptions of personal data, physical and laboratory examinations, clinical diagnoses, proposed treatments, and the outcomes of such treatments. Given a large database of such records in a medical specialty, a physician may query the database as to decisions and events that appear analogous to those involving the present patient, so as to display the collective, real-world experience bearing on the situation. In contrast to rule-based expert systems, which are (ideally) intended to replace a human expert with a machine, knowledge bases containing descriptions of actual problem events may be used only as decision-aiding tools. They are attractive, however, because their development is usually a by-product of organizational information systems and because their usefulness (to practice, research, continuing education, and so forth) increases with the volume of expert experience they acquire.

Public information utilities.

Aside from the proliferation of organizational information systems, new types of teleprocessing systems became available for use by the public during the 1970s. With the proliferation of electronic databases, the then-new industry of "database vendors" began to make these resources available via on-line database search systems. Today this industry operates for public access and uses hundreds of document databases, some of them in full text; corporate and industry data and news; stock quotations; diverse statistics and time series; and catalogs of products and services.

Recent services of public information utilities include transaction-processing systems: brokerage services to place on-line stock, bond, and options orders; home banking to pay bills and transfer funds; travel planning and reservations; and on-line catalog shopping. Some of these services combine on-line information retrieval (from, say, merchandise catalogs) and transaction processing (placing orders). Many include such functions as electronic mail and teleconferencing.

IMPACT OF COMPUTER-BASED INFORMATION SYSTEMS ON SOCIETY

Preoccupation with information and knowledge as an individual, organizational, and societal resource is stronger today than at any other time in history. The volume of books printed in 16th-century Europe is estimated to have doubled approximately every seven years. Interestingly, the same growth rate has been calculated for global scientific and technical literature in the 20th century and for business documents in the United States in the 1980s. If these estimates are reasonably correct, the growth of recorded information is a historical phenomenon, not peculiar to modern times. The present, however, has several new dimensions relative to the information resource: modern information systems collect and generate information automatically; they provide rapid, high-resolution access to the corpora of information; and they manipulate information with previously unattainable versatility and efficiency.

The proliferation of automatic data-logging devices in scientific laboratories, hospitals, transportation, and many other areas has created a huge body of primary data for subsequent analysis. Machines even generate new information: original

musical scores are now produced by computers, as are graphics and video materials. Electronic professional workstations can be programmed to carry out any of a variety of functions. Some of those that handle word processing not only automatically look for spelling and punctuation errors but check grammar, diction, and style as well; they are able to suggest alternative word usage and rephrase sentences to improve their readability. Machines produce modified versions of recorded information and translate documents into other languages.

Modern information systems also bring new efficiency to the organization, retrieval, and dissemination of recorded information. The control of the world's information store has been truly revolutionized, revealing its diversity in hitherto unattainable detail. Information services provide mechanisms to locate documents nearly instantaneously and to copy and move many of them electronically. New digital storage technologies make it economical for some to obtain for personal possession those collections equivalent to the holdings of entire libraries and archives. Alternately, access to information resources on electronic networks permits the accumulation of highly individualized personal or corporate collections in analog or digital form or a combination of both.

As the imprint of technology expands, some of the fundamental concepts of the field, which often took centuries to evolve, are strained. For instance, information technology forces an extension of the traditional concept of the document as a fixed, printed object to include bodies of multimedia information. Because of their digital form, these objects are easy to manipulate; they are split into parts, recombined with others, reformatted from one medium to another, annotated in real time by people or machines, and readied for display in many different formats on various devices. Control of these "living" documents, which mimic human association and processing of ideas and are expected to become one of the most common units of the digital information universe, is but one of the challenges for the emerging virtual library of humankind.

An equally significant new dimension of modern information systems lies in their ability to manipulate information automatically. This capability is the result of

representing symbolic information in digital form. Computer-based information systems are able to perform calculations, analyses, classifications, and correlations at levels of complexity and efficiency far exceeding human capabilities. They can simulate the performance of logical and mathematical models of physical processes and situations under diverse conditions. Information systems also have begun to mimic human cognitive processes: deductive inference in expert systems, contextual analysis in natural-language processing, and analogical and intuitive reasoning in information retrieval. Powerful information-transforming technologies now available or under development - data/text to graphics, speech to printed text, one natural language to another - broaden the availability of information and enhance human problem-solving capabilities. Computer visualization is dramatically altering methods of data interpretation by scientists; geographic information systems help drivers of the latest automobiles navigate cities; and interactive applications of networked multimedia computers may, for some, replace newspapers, compete with commercial broadcast television, and give new dimensions to the future of education and training at all levels of society.

Information systems applications are motivated by a desire to augment the mental information-processing functions of humans or to find adequate substitutes for them. Their effects have already been felt prominently in three domains: the economy, the governance of society, and the milieu of individual existence.

Information Processing and Information Systems

Effects on the economy.

Information systems are a major tool for improving the cost-effectiveness of societal investments. In the realm of the economy, they may be expected to lead to higher productivity, particularly in the industrial and service sectors - in the former through automation of manufacturing and related processes, in the latter through computer-aided decision making, problem solving, administration, and support of clerical functions. Awareness that possession of information is tantamount to a competitive edge is stimulating the gathering of technical and economic intelligence at the corporate and national levels. Similarly, concern is mounting

over the safeguarding and husbanding of proprietary and strategic information within the confines of organizations as well as within national borders. Computer crime, a phrase denoting illegal and surreptitious attempts to invade data banks in order to steal or modify records, or to release over computer networks software (called a virus) that corrupts data and programs, has grown at an alarming rate since the development of computer communications. In worst-case scenarios, computer crime is capable of causing large-scale chaos in financial, military, transportation, municipal, and other systems and services, with attendant economic consequences.

The growing number of information-processing applications is altering the distribution of labour in national economies. The deployment of information systems has resulted in the dislocation of labour and has already had an appreciable effect on unemployment in the United States. That country's economic recession during the early 1990s saw thousands of middle-management jobs relinquished, most permanently. The growth of computer-based information systems encourages a change in the traditional hierarchical structure of management (see below). As heavy automation is reverting production facilities from the labour-intensive nations to industrialized countries, the competitive potential of some of these nations is also likely to suffer an economic setback, at least in the short run. Singapore, a city-state of some three million people, has become very prosperous as giant foreign electronic firms located their manufacturing facilities there. It is bracing against such an economic setback by seeking to become the world's most intensive user and provider of electronic information systems for public services, international commerce and banking, and communications.

Effects on governance and management.

For much of the history of humankind, formal organizations have been better equipped than the citizenry to take advantage of information: their record-keeping practices were more mature and efficient, they possessed better facilities and skills to collect and interpret information; and - with computational aids - they are now able to profit from the powerful analytical tools provided by information

technology. Possession of information is not, however, tantamount to higher-quality governance or management, particularly if such possession is unilateral. As the number of recent political and financial scandals in various countries documents, it also entails possibilities of error and misuse.

It is the democratization of information, a characteristic of the last decades of the 20th century, that portends a beneficial impact on the quality of human governance and management. The public information and communication utilities that propagate this trend not only render the concerned citizen's access to information more equitable, they also help to forge informal societal networks that counterbalance the power of formal organizations and increasingly focus their style of management on consulting with the well-informed and on conveying greater concern. The concept of the "electronic town hall," an issue debated in the U.S. presidential elections of 1992, encapsulates the ideal of participatory democracy.

An environment that encourages the use of information technology and systems fosters what might be termed high information maturity on the part of the populace, a prerequisite of participatory democracy. Equitable access to information by all citizenry - rich and poor, privileged and disadvantaged - is one of the poignant societal issues facing humankind in the 21st century.

Computer-based information systems also impact the structure and management styles of corporations. The matrix organization, a structure in which departments and employees communicate directly with other organizational units, is an increasingly popular alternative to the hierarchical structure. Loose organizational decentralization imitates the observed principle of nature and of social organization suggesting that a unit size of roughly 150 persons communicates optimally and requires minimal managerial overhead. As business expansion and mergers extend the authoritative reach of large corporations and as the use of standard techniques of electronic document interchange forges flexible networks of firms in most industrial domains, leadership by consensus replaces authoritarian management. Information sharing and communication are the principal factors bringing about

these changes, and information systems constitute the foundation that makes such sharing and communication effective.

Effects on the individual.

An overt impact of modern information systems concerns the individual's standard and style of living. Information systems affect the scope and quality of health care, make social services more equitable, enhance personal comfort, provide a greater measure of safety and mobility, and extend the variety of leisure forms at one's disposal. More subtly but equally important, they also affect the content and style of an individual's work and in so doing perturb the social and legal practices and conventions to which one is accustomed. New kinds of information products and media necessitate a redefinition of the legal conventions regulating the ownership of products of the human intellect. Moreover, massive data-collecting systems bring into sharp focus the elusive borderline between the common good and personal privacy, calling for the need to safeguard stored data against accidental or illegal access, disclosure, or misuse.

Individuals cannot ignore the impact of automation and information-processing systems on their skills and jobs. Information technology makes obsolete, in part or in entirety, many human functions: first mechanical and repetitive tasks were affected; now clerical and paraprofessional tasks are being automated; and eventually highly skilled and some professional functions will be made unnecessary. Individuals performing these functions face the probability of shorter periods of employment and the need to adapt or change their skills. As technologies, including information technology, grow more sophisticated, their learning curves stretch or the required skills become narrower; continuing training and education are likely to become a way of life for both employee and employer. Unlike the slow, gradual evolution of human labour in past generations, present-day changes are occurring rapidly and with little warning. Unless society members anticipate these effects and prepare to cope with them mentally and in practice, job dislocations and forced geographic relocations may prove traumatic for employees and their families.

The perhaps more fundamental issue of paramount long-term significance for society has to do with the well-being of the human spirit in an increasingly knowledge-intensive environment. In such an environment, knowledge is the principal and perhaps most valuable currency. The growing volume and the rate of obsolescence of knowledge compel the individual to live in the continuous presence of, and frequent interaction with, information resources and systems. Effective use of these resources and systems may be a modern definition of literacy, while the absence of such a skill may very well result in intellectual and possibly economic poverty and inequity. There is a real danger that humans, unwilling or incapable or not given access to information, may be relegated to an existence that falls short of the human potential.

ИНСТИТУТ ЭКОНОМИКИ И БИЗНЕСА

Banks and Banking

Introduction

The principal types of banking in the modern industrial world are commercial banking and central banking. A commercial banker is a dealer in money and in substitutes for money, such as checks or bills of exchange. The banker also provides a variety of financial services. The basis of the banking business is borrowing from individuals, firms, and occasionally governments - i.e., receiving "deposits" from them. With these resources and also with the bank's own capital, the banker makes loans or extends credit and also invests in securities. The banker makes profit by borrowing at one rate of interest and lending at a higher rate and by charging commissions for services rendered.

A bank must always have cash balances on hand in order to pay its depositors upon demand or when the amounts credited to them become due. It must also keep a proportion of its assets in forms that can readily be converted into cash. Only in this way can confidence in the banking system be maintained. Provided it honours its promises (e.g., to provide cash in exchange for deposit balances), a bank can create credit for use by its customers by issuing additional notes or by making new loans, which in their turn become new deposits. The amount of credit it extends may considerably exceed the sums available to it in cash. But a bank is able to do this only as long as the public believes the bank can and will honour its obligations, which are then accepted at face value and circulate as money. So long as they remain outstanding, these promises or obligations constitute claims against that bank and can be transferred by means of checks or other negotiable instruments from one party to another. These are the essentials of deposit banking as practiced throughout the world today, with the partial exception of socialist-type institutions.

Another type of banking is carried on by central banks, bankers to governments and "lenders of last resort" to commercial banks and other financial institutions. They are often responsible for formulating and implementing monetary and credit

policies, usually in cooperation with the government. In some cases - e.g., the U.S. Federal Reserve System - they have been established specifically to lead or regulate the banking system; in other cases - e.g., the Bank of England - they have come to perform these functions through a process of evolution.

Some institutions often called banks, such as finance companies, savings banks, investment banks, trust companies, and home-loan banks, do not perform the banking functions described above and are best classified as financial intermediaries. Their economic function is that of channelling savings from private individuals into the hands of those who will use them, in the form of loans for building purposes or for the purchase of capital assets. These financial intermediaries cannot, however, create money (i.e., credit) as the commercial banks do; they can lend no more than savers place with them.

The development of banking systems

Banking is of ancient origin, though little is known about it prior to the 13th century. Many of the early "banks" dealt primarily in coin and bullion, much of their business being money changing and the supplying of foreign and domestic coin of the correct weight and fineness. Another important early group of banking institutions was the merchant bankers, who dealt both in goods and in bills of exchange, providing for the remittance of money and payment of accounts at a distance but without shipping actual coin. Their business arose from the fact that many of these merchants traded internationally and held assets at different points along trade routes. For a certain consideration, a merchant stood prepared to accept instructions to pay money to a named party through one of his agents elsewhere; the amount of the bill of exchange would be debited by his agent to the account of the merchant banker, who would also hope to make an additional profit from exchanging one currency against another. Because there was a possibility of loss, any profit or gain was not subject to the medieval ban on usury. There were, moreover, techniques for concealing a loan by making foreign exchange available at a distance but deferring payment for it so that the interest charge could be camouflaged as a fluctuation in the exchange rate.

Another form of early banking activity was the acceptance of deposits. These might derive from the deposit of money or valuables for safekeeping or for purposes of transfer to another party; or, more straightforwardly, they might represent the deposit of money in a current account. A balance in a current account could also represent the proceeds of a loan that had been granted by the banker, perhaps based on an oral agreement between the parties (recorded in the banker's journal) whereby the customer would be allowed to overdraw his account.

English bankers in particular had by the 17th century begun to develop a deposit banking business, and the techniques they evolved were to prove influential elsewhere. The London goldsmiths kept money and valuables in safe custody for their customers. In addition, they dealt in bullion and foreign exchange, acquiring and sorting coin for profit. As a means of attracting coin for sorting, they were prepared to pay a rate of interest, and it was largely in this way that they began to supplant as deposit bankers their great rivals, the "money scriveners." The latter were notaries who had come to specialize in bringing together borrowers and lenders; they also accepted deposits.

It was found that when money was deposited by a number of people with a goldsmith or a scrivener a fund of deposits came to be maintained at a fairly steady level; over a period of time, deposits and withdrawals tended to balance. In any event, customers preferred to leave their surplus money with the goldsmith, keeping only enough for their everyday needs. The result was a fund of idle cash that could be lent out at interest to other parties.

About the same time, a practice grew up whereby a customer could arrange for the transfer of part of his credit balance to another party by addressing an order to the banker. This was the origin of the modern check. It was only a short step from making a loan in specie or coin to allowing customers to borrow by check: the amount borrowed would be debited to a loan account and credited to a current account against which checks could be drawn; or the customer would be allowed to overdraw his account up to a specified limit. In the first case, interest was charged on the full amount of the debit, and in the second the customer paid interest only

on the amount actually borrowed. A check was a claim against the bank, which had a corresponding claim against its customer.

Another way in which a bank could create claims against itself was by issuing bank notes. The amount actually issued depended on the banker's judgment of the possible demand for specie, and this depended in large part on public confidence in the bank itself. In London, goldsmith bankers were probably developing the use of the bank note about the same time as that of the check. (The first bank notes issued in Europe were by the Bank of Stockholm in 1661.) Some commercial banks are still permitted to issue their own notes, but in most countries this has become a prerogative of the central bank.

In Britain the check soon proved to be such a convenient means of payment that the public began to use checks for the larger part of their monetary transactions, reserving coin (and, later, notes) for small payments. As a result, banks began to grant their borrowers the right to draw checks much in excess of the amounts of cash actually held, in this way "creating money" - i.e., claims that were generally accepted as means of payment. Such money came to be known as "bank money" or "credit." Excluding bank notes, this money consisted of no more than figures in bank ledgers; it was acceptable because of the public's confidence in the ability of the bank to honour its liabilities when called upon to do so.

When a check is drawn and passes into the hands of another party in payment for goods or services, it is usually paid into another bank account. Assuming that the overdraft technique is employed, if the check has been drawn by a borrower, the mere act of drawing and passing the check will create a loan as soon as the check is paid by the borrower's banker. Since every loan so made tends to return to the banking system as a deposit, deposits will tend to increase for the system as a whole to about the same extent as loans. On the other hand, if the money lent has been debited to a loan account and the amount of the loan has been credited to the customer's current account, a deposit will have been created immediately.

One of the most important factors in the development of banking in England was the early legal recognition of the negotiability of credit instruments or bills of

exchange. The check was expressly defined as a bill of exchange. In continental Europe, on the other hand, limitations on the negotiability of an order of payment prevented the extension of deposit banking based on the check. Continental countries developed their own system, known as giro payments, whereby transfers were effected on the basis of written instructions to debit the account of the payer and to credit that of the payee.

The business of banking

The business of banking consists of borrowing and lending. As in other businesses, operations must be based on capital, but banks employ comparatively little of their own capital in relation to the total volume of their transactions. The purpose of capital and reserve accounts is primarily to provide an ultimate cover against losses on loans and investments. In the United States capital accounts also have a legal significance, since the laws limit the proportion of its capital a bank may lend to a single borrower. Similar arrangements exist elsewhere.

FUNCTIONS OF COMMERCIAL BANKS

The essential characteristics of the banking business may be described within the framework of a simplified balance sheet. A bank's main liabilities are its capital (including reserves and, often, subordinated debt) and deposits. The latter may be from domestic or foreign sources (corporations and firms, private individuals, other banks, and even governments). They may be repayable on demand (sight deposits or current accounts) or repayable only after the lapse of a period of time (time, term, or fixed deposits and, occasionally, savings deposits). A bank's assets include cash (which may be held in the form of credit balances with other banks, usually with a central bank but also, in varying degrees, with correspondent banks); liquid assets (money at call and short notice, day-to-day money, short-term government paper such as treasury bills and notes, and commercial bills of exchange, all of which can be converted readily into cash without risk of substantial loss); investments or securities (substantially medium-term and longer term government securities - sometimes including those of local authorities such as states, provinces, or municipalities - and, in certain countries, participations and

shares in industrial concerns); loans and advances made to customers of all kinds, though primarily to trade and industry (in an increasing number of countries, these include term loans and also mortgage loans); and, finally, the bank's premises, furniture, and fittings (written down, as a rule, to quite nominal figures).

All bank balance sheets must include an item that relates to contingent liabilities (e.g., bills of exchange "accepted" or endorsed by the bank), exactly balanced by an item on the other side of the balance sheet representing the customer's obligation to indemnify the bank (which may also be supported by a form of security taken by the bank over its customer's assets). Most banks of any size stand prepared to provide acceptance credits (also called bankers' acceptances); when a bank accepts a bill, it lends its name and reputation to the transaction in question and, in this way, ensures that the paper will be more readily discounted.

Deposits.

The bulk of the resources employed by a modern bank consists of borrowed money (that is, deposits), which is lent out as profitably as is consistent with safety. Insofar as an increase in deposits provides a bank with additional cash (which is an asset), the increase in cash supplements its loanable resources and permits a more than proportionate increase in its loans.

An increase in deposits may arise in two ways. (1) When a bank makes a loan, it may transfer the sum to a current account, thus directly creating a new deposit; or it may arrange a line of credit for the borrower upon which he will be permitted to draw checks, which, when deposited by third parties, likewise create new deposits. (2) An enlargement of government expenditure financed by the central bank may occasion a growth in deposits, since claims on the government that are equivalent to cash will be paid into the commercial banks as deposits. In the first instance, with the increase in bank deposits goes a related increase in the potential liability to pay out cash; in the second case, the increase in deposits with the commercial banks is accompanied by a corresponding increase in commercial bank holdings of money claims that are equivalent to cash.

Taking one bank in isolation, an increase in its loans may result in a direct increase in deposits. This may occur either as a result of a transfer to a current account (as above) or a transfer to another customer of the same bank. Once again, there is an increase in the potential liability to pay out cash. On the other hand, if there has been an increase in loans by another bank (including an increase in central bank loans to the government), this may give rise to increased deposits with the first bank, matched by a corresponding claim to cash (or its equivalent). For these reasons a bank can generally expect that, if there is an increase in deposits, there will also be some net acquisition of cash or of claims for receipt of cash. It is in this way that an increase in deposits usually provides the basis for further bank lending.

Except in countries where banks are small and insecure, banks as a whole can usually depend on their current account debits being largely offset by credits to current accounts, though from time to time an individual bank may experience marked fluctuations in its deposit totals, and all banks in a country may be subject to seasonal variations. Even when deposits are repayable on demand, there is usually a degree of inertia in the deposit structure that prevents sharp fluctuations; if money is accepted contractually for a fixed term or if notice must be given before its repayment, this inertia will be greater. On the other hand, if a significant proportion of total deposits derives from foreign sources, there is likely to be an element of volatility arising from international conditions.

In banking, confidence on the part of the depositors is the true basis of stability. Confidence is steadier if there exists a central bank to act as a "lender of last resort." Another means of maintaining confidence employed in some countries is deposit insurance, which protects the small depositor against loss in the event of a bank failure. Such protection was the declared purpose of the "nationalization" of bank deposits in Argentina between 1946 and 1957; banks receiving deposits acted merely as agents of the government-owned and government-controlled central bank, all deposits being guaranteed by the state.

Reserves.

Since the banker undertakes to provide depositors with cash on demand or upon prior notice, it is necessary to hold a cash reserve and to maintain a "safe" ratio of cash to deposits. The safe ratio is determined largely through experience. It may be established by convention (as it was for many years in England) or by statute (as in the United States and elsewhere). If a minimum cash ratio is required by law, a portion of a bank's assets is in effect frozen and not available to meet sudden demands for cash from the bank's customers. In order to provide more flexibility, required ratios are frequently based on the average of cash holdings over a specified period, such as a week or a month. In addition to holding part of the bank's assets in cash, a banker will hold a proportion of the remainder in assets that can quickly be converted into cash without significant loss. No banker can safely ignore the necessity of maintaining adequate reserves of liquid assets; some prefer to limit the sum of loans and investments to a certain percentage of deposits, not allowing their loan-deposit ratio to run for any length of time at too high a level.

Unless a bank held cash covering 100 percent of its demand deposits, it could not meet the claims of depositors if they were all to exercise in full and at the same time their rights to demand cash. If that were a common phenomenon, deposit banking could not long survive. For the most part, the public is prepared to leave its surplus funds on deposit with the banks, confident that they will be repaid if needed. But there may be times when unexpected demands for cash exceed what might reasonably have been anticipated; therefore, a bank must not only hold part of its assets in cash but also must keep a proportion of the remainder in assets that can be quickly converted into cash without significant loss. Indeed, in theory, even its less liquid assets should be self-liquidating within a reasonable time.

A bank may mobilize its assets in several ways. It may demand repayment of loans, immediately or at short notice; it may sell securities; or it may borrow from the central bank, using paper representing investments or loans as security. Banks do not precipitately call in loans or sell marketable assets, because this would disrupt the delicate debtor-creditor relationships and increase any loss of

confidence, probably resulting in a run on the banks. Ready cash may be obtainable in this way only at a very high price. Banks must either maintain their cash reserves and other liquid assets at a high level or have access to a "lender of last resort," such as a central bank, able and willing to provide cash against the security of eligible assets. In a number of countries, the commercial banks have at times been required to maintain a minimum liquid assets ratio. But central banks impose such requirements less as a means of maintaining appropriate levels of commercial bank liquidity than as a technique for influencing directly the lending potential of the banks.

Among the assets of commercial banks, investments are less liquid than money-market assets such as call money and treasury bills. By maintaining an appropriate spread of maturities, however, it is possible to ensure that a proportion of a bank's investments is regularly approaching redemption, thereby producing a steady flow of liquidity and in that way constituting a secondary liquid assets reserve. Some banks, particularly in the United States and Canada, have at times favoured the "dumbbell" distribution of maturities, a significant proportion of the total portfolio being held in long-dated maturities with a high yield, a small proportion in the middle ranges, and another significant proportion in short-dated maturities. Following redemption, the banks usually reinvest all or most of the proceeds in longer-term maturities that in due course become increasingly short-term. Interest-rate expectations frequently modify the shape of a maturity distribution, and, in times of great uncertainty with regard to interest rates, banks will tend to hold the bulk of their securities at short term, and something like a T-distribution may then be preferred (mainly shorts, supported by small amounts of medium to longer dated paper). Investments and money-market assets merge into each other. The dividing line is arbitrary, but there is an essential difference: the liquidity of investments depends primarily on marketability (though sometimes it also depends on the readiness of the government or its agent to exchange its own securities for cash); the liquidity of money-market assets, on the other hand, depends partly on marketability but mainly on the willingness of the central bank to purchase them or

accept them as collateral for a loan. This is why money-market assets are more liquid than investments.

INDUSTRIAL FINANCE

Long-term and medium-term lending.

Banks that do a great deal of long-term lending to industry must ensure their liquidity by maintaining relatively large capital funds and a relatively high proportion of long-term borrowings (e.g., time deposits, or issues of bonds or debentures), as well as valuing their investments very conservatively. Such banks, notably the French banques d'affaires and the German commercial banks, have developed special means of reducing their degree of risk. Every investment is preceded by a thorough technical and financial investigation. The initial advance may be an interim credit, later converted into a participation. Only when market conditions are favourable is the original investment converted into marketable securities, and an issue of shares to the public is arranged. One function of these banks is to nurse an investment along until the venture is well established. Even assuming its ultimate success, a bank may be obliged to hold such shares for long periods before being able to liquidate them. In addition, they often retain an interest in a firm as an ordinary investment as well as to ensure a degree of continuing control over it.

The long-term provision of industrial finance in Britain and the Commonwealth countries is usually handled by specialist institutions, with the commercial banks providing only part of the necessary capital. In Japan the long-term financial needs of industry are met partly by special industrial banks (which also issue debentures as well as accepting deposits) and partly by the ordinary commercial banks. In Germany the commercial banks customarily handle long-term finance.

Since World War II the commercial banks in the United States have developed the so-called term loan, especially for financing industrial capital requirements. The attempt to popularize the term loan began in the economic depression of the 1930s, when the banks tried to expand their business by offering finance for a period of years. Most term loans have an effective maturity of little more than five years,

though some run for 10 years or more. They are usually arranged between the customer and a group of lending banks, sometimes in cooperation with other institutions such as insurance companies, and are normally subject to a formal term loan agreement. Banks in Britain, western Europe, the Commonwealth, and Japan began during the 1960s to give term loans both to industry and to agriculture.

Short-term lending.

Short-term loans are the core of the banking business even in countries where commercial banks make long-term loans to industry. Much short-term lending consists in the provision of working capital, but the banks also provide temporary finance for fixed capital development, aiding a customer until long-term finance can be found elsewhere.

Much of this short-term lending is done by overdraft, particularly in the United Kingdom and a number of the Commonwealth countries, or by way of "current account lending" in many western European countries. The overdraft permits a depositor to overdraw an account up to an agreed limit. In theory, overdrafts are repayable on demand or after reasonable notice has been given, but often they are allowed to run on indefinitely, subject to a periodic review. An advance is reduced or repaid whenever the account is credited with deposits and recreated when new checks are drawn upon it, interest being paid only on the amount outstanding.

An alternative method of short-term lending is to debit a loan account with the amount borrowed, crediting the proceeds to a current account; interest is usually payable on the whole amount of the loan, which normally is for a fixed period of time. (In Britain arrangements are sometimes more flexible, and the term of the loan may be set by oral agreement.)

In a number of countries, including the United States, the United Kingdom, France, Germany, and Japan, short-term finance is often made available on the basis of discountable paper - commercial bills or promissory notes. Some of this paper is usually rediscountable at the central bank, thus becoming virtually a liquid asset, unlike a bank advance or loan.

Credit may be offered with or without formal security, depending on the reputation and financial strength of the borrower. In many countries, a customer may use a number of banks, and these institutions usually freely exchange information about joint credit risks. In Britain and The Netherlands, however, most concerns tend to use a single banking institution for most of their needs.

Traditionally bankers took the view that the liabilities of a bank (in particular, its deposits) were more or less stable and concerned themselves primarily with the investment of these funds. Since the late 1950s and '60s, especially in North America and latterly in the United Kingdom, there has been a change in emphasis. Banks began to find it more difficult to obtain deposits. Interest rates rose to high levels, and banks were obliged to compete with each other and with other institutions for funds. At the same time, there was little point in paying a high rate of interest for money unless it could be employed profitably. Bankers began to relate the cost of borrowed money directly to the return on loans and investments. Previously the main limitation on a bank's expansion had been its ability to find profitable new business, but now the determining factor became the availability of funds to lend out. The essence of assets and liabilities management, as it came to be called, was deciding what kinds of new money to buy and what to pay for it. In the United States the liabilities side of bank balance sheets now included, inter alia, in much larger proportion than during the 1960s, repurchase agreements (under which securities are sold subject to an agreement to repurchase at a stated date), federal funds purchases (on the assets side, federal funds sales), excess balances of commercial banks and other depository institutions (regularly traded throughout the United States), negotiable certificates of deposit (which can be traded on a secondary market), and, for the larger banks, Eurocurrency borrowings, mostly Eurodollars (dollar balances held abroad). In the United Kingdom, "bought" money consisted of wholesale (i.e., large) deposits (on which money market rates were paid), negotiable certificates of deposit, interbank borrowings, and Eurocurrency purchases. This bought money could then be used to finance the loan demand, including term loans, long favoured in the United States but a more recent

innovation in the United Kingdom and elsewhere, where they were developed considerably in the 1970s. Although much of the lending financed by bought money was by way of term loans, these could be "rolled over," with an interest rate adjustment, every three or six months, and there could therefore be a measure of interest-rate matching and also sometimes a matching of maturities. In less sophisticated environments than North America and the United Kingdom, there was again an increasing emphasis on bought money to meet any expansion in loan demands (much of which was now term lending), with an adjustment at the margin when more funds were needed - e.g., wholesale deposits, certificates of deposit, interbank borrowings, and purchases of Eurocurrencies.

The principles of central banking

The principles of central banking grew up in response to the recurrent British financial crises of the 19th century and were later adopted in other countries. Modern market economies are subject to frequent fluctuations in output and employment. Although the causes of these fluctuations are various, there is general agreement that the ability of banks to create new money may exacerbate them. Although an individual bank may be cautious enough in maintaining its own liquidity position, the expansion or contraction of the money supply to which it contributes may be excessive. This raises the need for a disinterested outside authority able to view economic and financial developments objectively and to exert some measure of control over the activities of the banks. A central bank should also be capable of acting to offset forces originating outside the economy, although this is much more difficult.

RESPONSIBILITIES OF CENTRAL BANKS

The first concern of a central bank is the maintenance of a soundly based commercial banking structure. While this concern has grown to comprehend the operations of all financial institutions, including the several groups of nonbank financial intermediaries, the commercial banks remain the core of the banking system. A central bank must also cooperate closely with the national government.

Indeed, most governments and central banks have become intimately associated in the formulation of policy.

Relationships with commercial banks.

One source of economic instability is the supply of money. Even in relatively well-controlled banking systems, banks have sometimes expanded credit to such an extent that inflationary pressures developed. Such an overexpansion in bank lending would be followed almost inevitably by a period of undue caution in the making of loans. Frequently the turning point was associated with a financial crisis, and bank failures were not uncommon. Even today, failures occur from time to time. Such crises in the past often threatened the existence of financial institutions that were essentially sound, and the authorities sometimes intervened to prevent complete collapse.

The willingness of a central bank to offer support to the commercial banks and other financial institutions in time of crisis was greatly encouraged by the gradual disappearance of weaker institutions and a general improvement in bank management. The dangers of excessive lending came to be more fully appreciated, and the banks also became more experienced in the evaluation of risks. In some cases, the central bank itself has gone out of its way to educate commercial banks in the canons of sound finance. In the United States the Federal Reserve System examines the books of the commercial banks and carries on a range of frankly educational activities. In other countries, such as India and Pakistan, central banks have also set up departments to maintain a regular scrutiny of commercial bank operations.

The most obvious danger to the banks is a sudden and overwhelming run on their cash resources in consequence of their liability to depositors to pay on demand. In the ordinary course of business, the demand for cash is fairly constant or subject to seasonal fluctuations that can be foreseen. It has become the responsibility of the central bank to protect banks that have been honestly and competently managed from the consequences of a sudden and unexpected demand for cash. In other words, the central bank came to act as the "lender of last resort." To do this

effectively, it was necessary that the central bank be permitted either to buy the assets of commercial banks or to make advances against them. It was also necessary that the central bank have the power to issue money acceptable to bank depositors. But if a central bank was to play this role with respect to commercial banks, it was only reasonable that it or some related authority be allowed to exercise a degree of control over the way in which the banks conducted their business.

Most central banks now take a continuing day-to-day part in the operations of the banking system. The Bank of England, for example, has been increasingly in the market to ensure that the banks have a steady supply of cash, even during periods of credit restriction. It also lends regularly to the discount houses, supplementing their resources whenever the commercial banks feel the need to call back money they have on loan to them. In the United States the Federal Reserve System has operated in a similar way by buying and selling securities on the open market and by lending to dealers in government securities on the basis of repurchase agreements. The Federal Reserve may also discount paper submitted by the commercial banks through the Federal Reserve banks. The various techniques of credit control in use are discussed in greater detail below.

The evolution of those working relations among banks implies a community of outlook that in some countries is relatively recent. The whole concept of a central bank as responsible for the stability of the banking system presupposes mutual confidence and cooperation. For this reason, contact between the central bank and the commercial banks must be close and continuous. The latter must be encouraged to feel that the central bank will give careful consideration to their views on matters of common concern. Once the central bank has formulated its policy after a full consideration of the facts and of the views expressed, however, the commercial banks must be prepared to accept its leadership. Otherwise, the whole basis of central banking would be undermined.

The central bank and the national economy.

Relationships with other countries.

Since no modern economy is self-contained, central banks must give considerable attention to trading and financial relationships with other countries. If goods are bought abroad, there is a demand for foreign currency to pay for them. Alternatively, if goods are sold abroad, foreign currency is acquired that the seller ordinarily wishes to convert into the home currency. These two sets of transactions usually pass through the banking system, but there is no necessary reason why, over the short period, they should balance. Sometimes there is a surplus of purchases and sometimes a surplus of sales. Short-period disequilibrium is not likely to matter very much, but it is rather important that there be a tendency to balance over a longer period, since it is difficult for a country to continue indefinitely as a permanent borrower or to continue building up a command over goods and services that it does not exercise.

Short-period disequilibrium can be met very simply by diminishing or building up balances of foreign exchange. If a country has no balances to diminish, it may borrow, but normally it at least carries working balances. If the commercial banks find it unprofitable to hold such balances, the central bank is available to carry them; indeed, it may insist on concentrating the bulk of the country's foreign-exchange resources in its hands or in those of an associated agency.

Long-period equilibrium is more difficult to achieve. It may be approached in three different ways: price movements, exchange revaluation (appreciation or depreciation of the currency), or exchange controls.

Price levels may be influenced by expansion or contraction in the supply of bank credit. If the monetary authorities wish to stimulate imports, for example, they can induce a relative rise in home prices by encouraging an expansion of credit. If additional exports are necessary in order to achieve a more balanced position, the authorities can attempt to force down costs at home by operating to restrict credit.

The objective may be achieved more directly by revaluing a country's exchange rate. Depending on the circumstances, the rate may be appreciated or depreciated, or it may be allowed to "float."

Appreciation means that the home currency becomes more valuable in terms of the currencies of other countries and that exports consequently become more expensive for foreigners to buy. Depreciation involves a cheapening of the home currency, thus lowering the prices of export goods in the world's markets. In both cases, however, the effects are likely to be only temporary, and for this reason the authorities often prefer relative stability in exchange rates even at the cost of some fluctuation in internal prices.

Quite often governments have resorted to exchange controls (sometimes combined with import licensing) to allocate foreign exchange more or less directly in payment for specific imports. At times, a considerable apparatus has been assembled for this purpose, and, despite "leakages" of various kinds, the system has proved reasonably efficient in achieving balance on external payments account. Its chief disadvantage is that it interferes with normal market processes, thereby encouraging rigidities in the economy, reinforcing vested interests, and restricting the growth of world trade.

Whatever method is chosen, the process of adjustment is generally supervised by some central authority - the central bank or some institution closely associated with it - that can assemble the information necessary to ensure that the proper responses are made to changing conditions.

Economic fluctuations.

As noted above, monetary influences may be an important contributory factor in economic fluctuations. An expansion in bank credit makes possible, if it does not cause, the relative overexpansion of investment activity characteristic of a boom. Insofar as monetary policy can assist in mitigating the worst excesses of the boom, it is the responsibility of the central bank to regulate the amount of lending by banks and perhaps by other financial institutions as well. The central bank may even wish to influence in some degree the direction of lending as well as the amount.

An even greater responsibility of the central bank is that of taking measures to prevent or overcome a slump. Recessions, when they occur, are often in the nature of adjustments to eliminate the effects of previous overexpansion. Such adjustments are necessary to restore economic health, but at times they have tended to go too far; depressive factors have been reinforced by a general lack of confidence, and, once this has happened, it has proved extremely difficult to stimulate recovery. In these circumstances, prevention is likely to be far easier than cure. It has therefore become a recognized function of the central bank to take steps to preclude, if possible, any such general deterioration in economic activity.

For the central bank to be effective in regulating the volume and distribution of credit so that economic fluctuations may be damped, if not eliminated, it must at least be able to regulate commercial bank liquidity (the supply of cash and "near cash"), because this is the basis of bank lending. Monetary authorities in a number of countries have begun to resort increasingly to the management of monetary aggregates as a basic policy. This does not mean an uncritical acceptance of monetarist philosophy but rather what the U.S. economist and banker Paul A. Volcker has called "practical monetarism." In addition to the Federal Reserve in the United States, a growing number of western European countries have adopted the practice of setting growth targets for the money supply and sometimes other monetary targets as well (like domestic credit expansion), usually setting some range of allowable variation. Japan has had reservations and has preferred to indicate monetary projections or forecasts, partly because of the difficulty of changing a set target should it become necessary. Nor is there any great degree of consensus as to which target or aggregate to employ. In general terms, choice of a particular aggregate as a basis for reference would be linked to the theories - more or less explicit - on which the actions of a particular central bank are based and also on the state of the country's economy and its financial environment. Where there are publicly declared targets, these can have an important effect by the very fact of being announced.

There is now little dispute about the broad objectives, though the techniques of control are various and depend to some extent on environmental factors. It would be incorrect to suppose, however, that the actions of the central bank can, unaided, achieve a high degree of stability. It can by wise guidance contribute to that end, but monetary action is in no sense a panacea; at all times, the degree to which it is likely to be effective depends on the provision of an appropriate fiscal environment.

Banking services.

Another responsibility of the central bank is to ensure that banking services are adequately supplied to all members of the community that need them. Some areas of a country may be "under-banked" (e.g., the rural areas of India and the northern and more remote parts of Norway), and central banks have attempted, directly or indirectly, to meet such needs. In France, this need underlay the early extension of branches of the Bank of France to the départements. In India the authorities encouraged the opening of "pioneer" branches by the former Imperial Bank of India and its successor, the State Bank of India, latterly by all the nationalized banks, and particularly their extension to rural and semirural areas. In Pakistan, officials of the State Bank of Pakistan played an active part in the foundation of the semipublic National Bank of Pakistan with a similar objective in view.

A different sort of problem arises when the business methods of existing banks are unsatisfactory. In such circumstances, a system of bank inspection and audit organized by the central banking authorities (as in India and Pakistan) or of bank "examinations" (as in the United States) may be the appropriate answer. Alternatively, the supervision of bank operations may be handed over to a separate authority, such as France's Banking Control Commission or South Africa's Registrar of Banks.

In developing countries, central banks may encourage the establishment and growth of specialist institutions such as savings institutions and agricultural credit or industrial finance corporations. These serve to improve the mechanism for

tapping existing liquid resources and to supplement the flow of funds for investment in specific fields.

Responsibilities to the government.

Central banks have over the years acquired a number of well-defined responsibilities to their respective national governments. Some, notably the Bank of England, developed into central banks after being, in origin, bankers to the government. More recently it has become a matter of course for a new central bank to accept responsibility for the financial affairs of its government. The reasons are self-evident. Government transactions have become of increasing importance in influencing the workings of the economy, and the institution that holds the government's account is in a strategic position to cushion the commercial banks against the impact of large movements of cash originating in this way. As banker to the government, furthermore, the central bank has an obvious responsibility to provide routine banking services, such as arranging loan flotations and supervising their service, renewal, and redemption. The central bank also usually issues the currency.

Equally important are its responsibilities as an adviser on the probable monetary consequences of any proposed action. In this role the central bank should scrutinize the government's proposals with a certain amount of objectivity and state its point of view with vigour. One may cite a now-famous dictum of Montagu Norman as governor of the Bank of England:

I think it is of the utmost importance that the policy of the Bank and the policy of the Government should at all times be in harmony - in as complete harmony as possible. I look upon the Bank as having the unique right to offer advice and to press such advice even to the point of "nagging"; but always of course subject to the supreme authority of the Government.

Many central banks are now nationalized, reflecting the increasingly general recognition of the significance of the central bank's role as a servant, if not a creature, of the government. This development is also, in a way, a final recognition of the central bank as a responsible public institution whose major function is to

serve the community as a whole, untrammelled by narrow dictates of profit and loss. Most central banks, nevertheless, make very handsome profits.

TECHNIQUES OF CREDIT CONTROL

Central banks have developed a variety of techniques for influencing, regulating, and controlling the activities of commercial banks. These may be divided into (1) the so-called classical, or indirect, techniques and (2) various direct controls. The classical techniques make use of open-market purchases or sales by the central bank of certain types of assets that are invariably associated with fluctuations in interest rates. Direct, or quantitative, credit controls are employed to influence the cash and liquidity bases of commercial bank lending by means of freezing or unfreezing their liquid resources; sometimes ceilings are imposed on bank loans.

Open-market operations.

The way in which open-market operations influence the cash reserves and, through them, the general liquidity of the commercial banks is essentially simple. If the central bank buys securities in the open market, the cash it offers in exchange adds to the reserves of the banks; if the central bank sells securities in the open market, the cash necessary to pay for them is either withdrawn from the banks' reserves or obtained by diminishing holdings of other assets (with the possibility of capital losses in consequence of these sales). It does not matter whether this buying and selling takes place between the central bank and the commercial banks directly or between the central bank and other financial sectors, including the public at large, since these are the customers of the commercial banks.

Open-market operations are invariably associated with related changes in one or more "strategic" rates of interest, the most influential of these rates being the minimum rate at which the central bank does business (the bank rate, or the discount rate), since other rates tend to move in sympathy with it. The central bank seeks to achieve an appropriate and consistent structure of interest rates. If a particular rate structure is desired (e.g., prior to a new issue of government securities or in order to change the emphasis of institutional investment between, say, long-term and short-term securities), it may be necessary to precondition the

market by means of open-market operations. To achieve its purposes the central bank must possess (if it is selling) or be willing to absorb (if it is buying) the appropriate types of securities.

In London the specialist banks known as discount houses effectively put to work the revolving fund of cash that circulates through the British banking system. If temporarily there is an inadequate supply of cash, the Bank of England either lends on a short-term basis or buys some of the assets held by the discount market. (From 1980 there was a shift in emphasis from lending to open-market operations, especially by dealing in bankers' acceptances.) Alternatively, the Bank of England may buy assets from the clearing banks (the large joint-stock banks), which then make the relevant moneys available to the market. On the other hand, if the discount market is oversupplied with funds, the Bank of England sells treasury bills, in this way mopping up the excess of cash. These transactions are known as smoothing-out operations. In addition, the Bank of England is also responsible for managing the national debt, and, whether the object is to influence the flows of money or not, such transactions in fact have monetary effects.

In the United States the Federal Reserve System regulates the money supply. Within the Federal Reserve System, the Federal Open Market Committee is the most important single policy-making body. It is presided over by the chairman of the Board of Governors, with the president of the Federal Reserve Bank of New York as its permanent vice chairman. The main responsibility of the Open Market Committee is to decide upon the timing and amount of open-market purchases or sales of government securities. Since open-market operations must obviously be consistent with other aspects of monetary and credit policy, it is in the committee that broad agreement is reached on matters such as changes in discount rates or reserve requirements.

One of the big differences between London and New York is that the central banking authorities in New York maintain direct relationships more or less continuously with the nonbank government securities dealers as well as with the commercial banks. The Federal Reserve Bank of New York may make temporary

accommodation available to some 35 primary dealers (including certain banks) under a repurchase agreement, whereby securities are sold to the bank under an agreement that they be repurchased after a stipulated time. These agreements are made only for the purpose of supplying reserves to the banking system, but from the dealer's standpoint they are helpful in financing portfolios. Such repos, as they are called, may also be done with foreign official accounts. Since early 1966 the bank has also been prepared to mop up money by undertaking reverse repurchase agreements, in which the dealers act as intermediaries for large commercial banks with temporarily surplus money that they are prepared to place against bills, subject to the bank's repurchasing them a few days later; the commercial bank concerned lends the dealer the money to finance the holding of the bill. Similar arrangements are also made by the Federal Reserve directly with bank dealers.

All member banks of the Federal Reserve System, and now also other depository institutions, have direct access to the discount service of their Federal Reserve Bank, of which there is one in each of 12 districts. This is a privilege, however, and not a right. In the early years of the system, the banks would sell discountable paper to the Federal Reserve, but now they usually borrow against a pledge of government securities held in safe custody with the Federal Reserve Bank in question. The Federal Reserve lends for a number of purposes but always at a time of general stress. It is assumed that, as the pressure abates, borrowing banks will repay their indebtedness as quickly as possible. Under ordinary conditions, the continuous use of Federal Reserve credit by a member bank over a considerable period is not regarded as appropriate.

Direct control of assets.

The so-called classical techniques of credit control - open-market operations and discount policy - can be employed only where there is a sufficiently developed complex of markets in which to buy and sell assets of the type that commercial banks ordinarily hold. Direct credit controls have a wider range of application. They may be used either as a substitute for the classical techniques or as a supplement to them. Direct controls are more likely to be resorted to when the

money market is undeveloped, because then a central bank can only impose its authority by means of direct action. This is often the situation facing a newly established central bank. Rather than wait for the slow evolution of a money market, the authorities may provide the central bank from the start (as in Pakistan, the Philippines, Sri Lanka, and Malaysia) with very full powers to control the banking system.

The aim in imposing a direct, quantitative regulation of credit is to curb inflationary pressures that may result from an expansion of commercial bank lending. This can be done in four main ways: (1) the commercial banks may be required to maintain stated minimum reserve ratios of cash to deposits, a stated liquid assets ratio, or some combination of both; (2) part of the cash resources of the commercial banks may be immobilized at the discretion of the central bank; (3) ceilings may be imposed on the amount of accommodation to be made available to the commercial banks at the central bank (sometimes referred to as "discount quotas"); and (4) a ceiling may be prescribed for commercial bank lending itself.

Minimum reserve requirements.

The variation of minimum cash reserve requirements as a direct means of quantitative credit control has become increasingly general in recent years. The practice has largely derived from experience in the United States. In its origin the U.S. insistence on stated minimum reserve requirements for commercial banks was simply a means of prescribing minimum standards of sound behaviour. Only later did such ratios come to be seen as a useful supplementary quantitative credit control.

The power granted by the Banking Act of 1935 to the Federal Reserve System to determine the cash reserves of the commercial banks in the United States was employed for the first time during the boom of 1936-37, and periodic variation of minimum reserve requirements subsequently came to be recognized as an appropriate technique for controlling the money supply. The Federal Reserve Board's decisions were sometimes subject to considerable criticism, but, as it became more experienced in the use of this technique, variation in reserve

requirements combined with other measures came to be regarded as a useful means of cushioning the economy against a recession. The variation of reserve requirements did not prove as effective in preventing inflation, largely because of the government's insistence that the Federal Reserve simultaneously support the prices of government bonds through open-market operations. This insistence was abandoned by the Treasury in March 1951. Since then, much greater emphasis has been placed on the use of open-market operations, which had become more effective, and the importance of varying minimum reserve requirements as a means of controlling the credit base has diminished in the United States. The technique is still widely used, however, in many countries.

In some countries, the authorities require the maintenance of minimum liquid assets ratios. This is often combined with minimum requirements for cash reserves, as in India, Pakistan, and Germany, though not always (in France, for example, until 1967 there were no minimum cash reserve requirements). Where prescribed minima relate to liquid assets and not to cash as such, reserves are held in the form of earning assets - an important distinction from the point of view of the commercial banks.

An important step toward a uniform and explicit minimum liquidity ratio for the London clearing banks was taken in 1951, when the governor of the Bank of England indicated to the banks that a liquidity ratio of from 32 to 28 percent would be regarded as normal and that it would be undesirable for the ratio to be allowed to fall below 25 percent. By 1957 a fairly rigid 30-percent minimum was in place (it was reduced to 28 percent in 1963). After 1946 the London clearing banks (but not the Scottish banks) also observed a more or less fixed cash ratio of 8 percent. A new element was introduced in 1960, when the Bank of England launched its system of "special deposits" as a means of reinforcing other methods of credit control. Calls were made from time to time on the London clearing banks to deposit with the Bank of England by a specified date some specified percentage of their gross deposits; similar arrangements applied to the Scottish banks, but the calls were smaller. This system lasted until 1971, when a new 12.5-percent

minimum reserve ratio (excluding till cash) was introduced. This ratio related to "eligible liabilities" (primarily sterling deposits of up to two years maturity, including sterling certificates of deposit). The banks could also be required to place special deposits with the Bank of England. These arrangements were replaced in August 1981 by a voluntary holding of operational funds with the Bank of England by the London clearing banks ("for clearing purposes") and a uniform requirement of 0.5 percent of an institution's eligible liabilities that would be applied to all banks and licensed deposit-takers with eligible liabilities averaging more than 10,000,000. All banks that were eligible acceptors were also normally required to hold an average equivalent to 6 percent of their eligible liabilities either as secured money with discount houses or as secured call money with money brokers and gilt-edged jobbers, but the amount held in the form of secured money with a discount house was not normally to fall below 4 percent of eligible liabilities. This money became known as "club money."

The use of variable minimum reserve requirements as a means of credit control can, if carried far enough, produce results, especially when the requirements include the holding of cash balances. It is more useful as an anti-inflationary weapon than as a means of countering recession, since it cannot overcome a possible unwillingness of the banks to lend or of their customers to borrow. It is a somewhat clumsy technique, however, and cannot make adequate allowance for the special needs of different institutions.

Immobilization of cash resources.

A second group of direct quantitative credit controls involves keeping a portion of the cash resources of commercial banks immobilized at the discretion of the central bank. Two leading examples of this technique were the use of the Treasury Deposit Receipt (TDR) in the United Kingdom during and after World War II and the "special account procedure" adopted in Australia in 1941. Both were means of immobilizing the increased liquidity deriving from wartime government expenditure.

The direct issue of Treasury Deposit Receipts at a nominal rate of interest to banks in the United Kingdom began in July 1940. They were not negotiable in the market nor transferrable between banks, but they could be tendered in payment for government bonds (and tax certificates); hence, during the war years they had a limited degree of liquidity. The Bank of England communicated to the banks collectively the amount of the weekly call, which was divided among them in proportion to their deposits. After the war, TDR's were replaced by treasury bills; in order to reduce the consequent high liquidity of the banks, there was a "forced funding" of 1,000,000,000 of treasury bills in November 1951, which were required to be invested in Serial Funding Stocks.

The special account procedure introduced in Australia in 1941 had a similar objective. The surplus investable funds of the Australian trading banks, defined as the amount by which each bank's total assets in Australia at any time exceeded the average of its total assets in Australia in August 1939, were required to be placed in special deposit accounts with the Commonwealth Bank (then the central bank) at a nominal rate of interest. A bank was not to withdraw any sum from its special account except with the consent of the Commonwealth Bank; during the war years, the bank generally directed the trading banks to lodge in their special accounts each month an amount equal to the increase in their total assets in Australia during the preceding month, although as a rule a lodgment was not required if it was known that a rise in assets would be followed by an early fall. Legislation in 1945 adopted the special account procedures as a means of general credit control (e.g., to curb inflation), but the provisions were made more flexible. In 1953 a more complicated formula was introduced, and in 1960 the system was abandoned in favour of minimum reserve ratios.

Direct control of loans.

Accommodation ceilings.

Some countries have tried limiting the amount of accommodation that the central bank may make available to the commercial banks. The difficulty in this type of quantitative credit control is to make it effective while also allowing for changes in

the economy; its most obvious use is as a means of checking inflation, but, if the upward pressures on prices are strong, there is a temptation to increase the ceilings so that the restraint then becomes little more than a temporary check.

Usually, it is only when a control begins to be felt and to affect bank profits that the banks become really sensitive to changes in credit policy and the implementation of the control becomes truly effective. The postwar experience of France is a case in point. Plafonds, or "ceilings," were first introduced in France in 1948. Rediscount ceilings (or discount quotas) were fixed for each bank, though some categories of paper were excluded. Ceilings could be increased or (after 1957) reduced.

From the authorities' point of view, the chief difficulty in operating this control was the persistent building up of pressure against the ceilings. This was met partly by upward revisions in the ceilings themselves and partly by instituting a number of safety valves. The degree of elasticity required constituted the chief weakness of the ceiling technique. The central bank was constantly under pressure to adjust the ceilings upward. Some upward revisions were unavoidable, but the problem was to decide which claims were legitimate and which not. Much bilateral bargaining took place between the Bank of France and individual commercial banks, but the banks continued to complain that the strictness of the control was excessive and that the technique was lacking in flexibility.

The inadequacies of the plafonds technique in its original form became apparent when prices began to rise rapidly during the Korean War boom, and even the built-in safety valves failed fully to accommodate the pressures on bank liquidity. The need to strengthen the mechanism was obvious, and this was attempted in 1951. Previously, rediscounts had frequently exceeded the ceilings during the month and were only brought within the plafonds by special action (e.g., through open-market purchases). The situation was brought under control by introducing a secondary ceiling to which a penalty rate of interest was applied. This was extended in 1958 to permit rediscounts even beyond the secondary ceiling, provided a further penalty was paid; each application, however, was scrutinized by the Bank of

France. The system lasted until about the spring of 1964, though it did not finally disappear until 1968, when it was largely replaced by Bank of France operations in the open market. After early 1967, banks also were subject to minimum reserve requirements.

Plafonds, or discount quotas, also are employed in Germany. They were introduced in West Germany in 1952 and strengthened in 1955. Quotas may be reduced periodically (after 1964 they were also used to discourage institutions from borrowing abroad). Again there were safety valves (although less generous than in France) and the possibility of extra accommodation (Lombard credits) at a higher rate. In some circumstances, supplementary quotas might be approved for up to six months. A bank might also raise funds through the money market, though likely at higher cost. Discount quotas are still an important tool of credit control in Germany.

Other countries have employed this technique, including Sweden, where for a time the central bank imposed formal or informal ceilings on banks and sometimes on finance companies. If the banks failed to observe the ceiling, a penalty was applied based on the amount of the excess borrowing and its duration. In Finland, commercial banks have at times been able to borrow limited amounts from the Bank of Finland by way of traditional credit quotas. Beyond these quotas, funds could formerly be obtained as supra-quota credit at a higher rate, but banks now are forced into the official call-money market. Denmark, too, has permitted borrowing from the central bank in tranches, with higher (penalty) rates applying after the first tranche of the loan quota has been resorted to, a practice that can be expensive.

General ceilings on credit.

Attempts have been made to prescribe a general ceiling within which the quantity of commercial bank lending must be held. This is even more difficult to achieve. One example of such an attempt was the adoption of a "rising ceiling" by Chile in 1953. All banks were required not to expand the volume of their loans to businesses and individuals by more than 1.5 per-cent a month, using as their basis

the average of a bank's advances on selected dates in 1953. Certain types of loans were forbidden, and bank resources were to be directed to productive and distributive activities that really contributed to the expansion of the national economy. Banks were also required to provide information on the destination of their loans. In succeeding years, adjustments were made on several occasions in the maximum permitted credit increase, expressed either as a percentage of advances or sometimes as a total for the banking system as a whole. In 1959 all quantitative credit restrictions were removed, and banks were permitted to advance funds up to their financial capacity, provided that they operated within the general banking law. There was no evidence the controls had been effective, but the major problem in Chile was budgetary rather than monetary. A temporary ceiling on loans was imposed by agreement in Canada (in 1951-52), The Netherlands (1957-58), and France (1958-59).

The United Kingdom had considerable experience with this type of ceiling, introducing it as a temporary measure in 1955, when the banks were asked to bring their advances down by an average of 10 percent. Later an attempt was made to impose a true ceiling, requiring that bank advances not exceed the average of the period October 1956 to September 1957. This was continued until July 1958. Again, in 1961, the authorities indicated the banks must aim at checking the rate of rise in bank advances; this came to be interpreted as a request that the level of advances at the end of 1961 be no higher than in the previous June. The banks also were not to encourage an increase in the volume of commercial bills. The request was modified in May 1962 and largely withdrawn in October; but it was made again in May 1965, when the clearing banks were requested not to increase their advances to the private sector, at an annual rate of more than about 5 percent, in the 12 months to mid-March 1966 (likewise with commercial bills). Other financial institutions were requested to observe a comparable degree of restraint. For 12 months after March 1966, advances and discounts, allowing for seasonal factors, were not permitted to rise above levels set for March 1966. This represented an intensification of the credit squeeze because prices were rising. The

credit restriction led to a falling off in business confidence, and, consequently, toward the end of 1966, bank lending was well below the official ceiling. In April 1967, authorities announced a change in techniques, with an emphasis on making calls to special deposits, but the ceilings returned again in November 1967. There was to be no increase in bank advances to the private sector (excluding exports and shipbuilding) except for seasonal reasons. In May 1968 a new ceiling was instituted for all such lending (including that for exports and shipbuilding); the clearing banks were asked to restrict the total of this lending, after seasonal adjustment, to 104 percent of the November 1967 figure, with priority to be given to finance for exports and for activities directly related to improving the balance of payments. The restrictions also extended to other types of credit. Credit became even tighter (in March 1969) when the ceiling was reduced to 98 percent of the November 1967 level. The banks had considerable difficulty in meeting this requirement and agreed merely to "do their best." Advances increased above the ceiling, and, as a penalty, the interest paid by the Bank of England on special deposits was halved. Not until late 1969 did it become clear that the authorities were prepared to abandon their long campaign to get bank loans down to the target figure. The ceiling was subsequently replaced by minimum reserve requirements. The system of quantitative credit control requires, for its successful implementation, the full cooperation of the banking community. In the United Kingdom, where banks base much of their lending on the overdraft technique, the system was very unpopular.

In France, however, the encadrement du crédit, as it is called, temporarily imposed in 1958-59, was revived during the first half of 1973. Subject to certain exclusions (e.g., certain investment credits, agricultural credits, export credits, the financing of energy savings and innovation, leasing transactions, and special medium-term construction loans), the mechanism chosen was to permit a certain percentage rate of growth in bank credits in relation to a particular month in the previous year, these limits being fixed quarterly and subject to variation from time to time. Subsequently, in early 1975, reference was made to a fixed base defined as equal

to an index of 100, in relation to which the index might be increased (or decreased) and credit expanded (or contracted). The system was further refined to vary the rate of change of credits within different financial sectors, and it has been subject in the interests of flexibility to many amendments over the years. In effect, there has been a combination of quantitative and qualitative credit controls.

In addition to regulating the quantity of credit, central banks have sometimes attempted to influence the directions in which the commercial banks lend. A loose system of control prevailed in the United Kingdom during World War II and afterward, based initially on directives from the Capital Issues Committee and later on requests from the Bank of England. A highly formalized technique was employed in Australia during the war and earlier postwar years; detailed and specific instructions were given to the trading banks, marginal cases being referred to the central bank. The system of Voluntary Credit Restraint in the United States in 1951 was similar. The more formal controls seemed to be no more effective than the looser system employed in the United Kingdom.

Selective controls have been imposed on consumer installment finance in the United States and elsewhere (e.g., by stipulating the percentage of deposit that is required and the length of the term over which repayments may be made). Even when these are not varied in order to serve as a control over credit, there is a case for insisting on such requirements for prudential reasons. In the United States, under the Securities Exchange Act of 1934, the Federal Reserve can vary the margins that purchasers of securities must pay in cash, thereby limiting the credit available for this purpose. The structure of modern banking systems

The banking systems of the world have many similarities, but they also differ, sometimes in quite material respects. The principal differences are in the details of organization and technique. The differences are gradually becoming less pronounced because of the growing efficiency of international communication and the tendency in each country to emulate practices that have been successful elsewhere.

Banking systems may be classified in terms of their structure as unit banking, branch banking, or hybrids of the two. For example, unit banking prevails in large areas of the United States. In other countries it is more usual to find a small number of large commercial banks, each operating a highly developed network of branches. This is the system used in England and Wales, among others. Examples of hybrid systems include those of France, Germany, and India, where banks that are national in scope are supplemented by regional or local banks. Some of these hybrid systems are slowly changing their character, the banks becoming fewer in number and individually larger, with a larger number of branches.

UNIT BANKING: THE UNITED STATES

Bank organization in the United States during the years after World War II was still passing through a phase of structural development that many other countries had completed some decades earlier. Development in the United States has been subject to constraints not found elsewhere. The federal Constitution permits both the national and state governments to regulate banking. Some states prohibit branch banking, largely because of the political influence of small local bankers, thus encouraging the establishment and retention of a large number of unit banks.

Even in its early years, the United States had an unusually large number of banks. As the frontiers of settlement were pushed rapidly westward, banks sprang up across the country. One reason for this was the demand for capital in the expanding frontier economy. There was also an obvious need for a large number of banks to serve the diverse and rapidly expanding demands of a growing and constantly migrating population. It must be remembered, too, that at this time communications between the frontiers of settlement and the established centres of commerce and finance were still inadequately developed.

As long as communications remained imperfect, the existence of large numbers of competing institutions is not difficult to explain. The subsequent failure of bank mergers or amalgamations to produce a concentration of financial resources in the hands of large banking units can be attributed in part to the character of the federal Constitution as noted above. Among the people, moreover, there was a widespread

distrust of monopoly and a deep-rooted fear that a "money trust" might develop. This went hand in hand with a political philosophy that emphasized the virtues of individualism and free competition; restrictions on branching, merging, and on the formation of holding companies were a feature of both the state and the federal banking laws. Where permitted, however, bank branches are numerous in the United States (especially in California and in New York); in states in which branching is prohibited, one often finds local bank monopolies in small towns. Interstate banking is prohibited by federal law, but large banking organizations have provided financial services (e.g., through loan offices and offices of nonbank subsidiaries) for many years across state lines. A number of states have passed limited interstate or reciprocal banking laws, so that banks in other states with similar laws can acquire or merge with local banks. The banking system of the United States would not work without a network of correspondent bank relationships, which are more highly developed there than in any other country.

From the 1970s there was an acceleration in the evolution of U.S. banking patterns. Unregulated financial institutions (and some nonfinancial institutions) moved into traditional banking activities; at the same time, depository institutions began offering a fuller range of financial services. Money-market mutual funds, for example, secured access to open-market interest rates for investors with relatively small amounts of money. Securities firms and insurance companies moved aggressively into providing a range of liquid financial instruments. Likewise, large manufacturing and retail firms moved into the commercial and retail lending businesses - e.g., by acquiring a savings and loan association, a securities brokerage house, an industrial loan company, a consumer banking business, or even a commercial bank. Meanwhile, depository institutions developed a number of new services, most notably the Negotiable Order of Withdrawal (NOW) account, an interest-bearing savings account with a near substitute for checks. These appeared first in 1972 in New England and after 1980 spread to the whole nation; they were offered both by commercial banks and by thrift institutions. Share drafts at credit unions also became a means of payment, and after 1978 the

automatic transfer services of commercial banks permitted savings account funds to be transferred automatically to cover overdrafts in checking accounts. So-called Super-NOW accounts (with no interest rate ceilings and unlimited checking facilities with a minimum balance) were subsequently introduced, along with money-market deposit accounts, free of interest rate restrictions but with limited checking.

Rapid changes in financial structure and the supply of financial services posed a host of questions for regulators, and, after much discussion, the Depository Institutions Deregulation and Monetary Control Act was passed in 1980. The object was to change some of the rules - many of them obsolete - under which U.S. financial institutions had operated for nearly half a century. The principal objectives were to improve monetary control and equalize more nearly its cost among depository institutions; to remove impediments to competition for funds by depository institutions, while allowing the small saver a market rate of return; and to expand the availability of financial services to the public and reduce competitive inequalities among financial institutions offering them. The major changes were: (1) Uniform Federal Reserve requirements were phased in on transaction accounts (demand deposits, NOW accounts, telephone transfers, automatic transfers, and share drafts) at all depository institutions - commercial banks (whether Federal Reserve members or not), savings and loan associations, mutual savings banks, and credit unions. (2) The Federal Reserve Board was authorized to collect all data necessary for the monitoring and control of money and credit aggregates. (3) Access to the discount window at Federal Reserve banks was widened to include any depository institution issuing transaction accounts or nonpersonal time deposits. (4) The Federal Reserve was to price its services, to which all depository institutions would now have access. (5) Regulation Q, which had long set interest-rate ceilings on deposits, was to be phased out over a six-year period. (6) An attempt was made to grasp the nettle of the state usury laws. (7) NOW accounts were authorized on a nationwide basis and could be offered by all depository institutions. Other services were extended. (8) The permissible activities of thrift

institutions were broadened considerably. (9) Deposit insurance at commercial banks, savings banks, savings and loan associations, and credit unions was raised from $40,000 to $100,000. (10) The "truth in lending" disclosure and financial regulations were simplified to make it easier for creditors to comply.

BRANCH BANKING: THE UNITED KINGDOM

If the United States banks can be taken as representative of a unit banking system, the British system is the prototype of branch banking. Its development was linked to the growth of transportation and communications, for otherwise banks cannot clear checks drawn on other banks and effect remittances speedily and efficiently. The Scots favoured branch banking from the very beginning (the Bank of Scotland was founded in 1695), but at first they were not very successful - largely because of poor communications and the difficulty of supplying branches with adequate amounts of coin. Not until after the Napoleonic Wars, when the road system of Scotland had been greatly improved, did branch banking begin to develop vigorously there. As the Industrial Revolution progressed and as the size of businesses increased, the structure of English banking underwent a corresponding change. Greater resources were required for lending, and banks also needed more extensive interconnections in order to provide an increasing range of services. Where banks remained small, they were frequently unable to take the strain of the larger demand; they tended to become overextended and often failed.

The growth in size of banks was also greatly encouraged by legislation that encouraged joint-stock ownership, beginning in 1826. Joint-stock ownership, which reduced the risk to any individual, must be distinguished from limited liability, which did not become widely accepted until the failure of the City of Glasgow Bank in 1878 demonstrated the need for a legal device to protect the stockholder. The early joint-stock banks tended to remain localized in their business interests; it was only gradually (with the spread of limited liability and disclosure of accounts) that amalgamations began to convert the banking system in England and Wales into its highly concentrated modern form. The main movement was completed before World War I, though there was to be a further degree of

concentration in the years after World War II. By these means, British banks were able to attract deposits from all parts of the country and to spread the banking risk over a wide range of industries and areas.

HYBRID SYSTEMS

A third group of banking systems differs from the unit banking system of the United States and also from the branch banking systems of countries that have followed the British model (such as Australia, Canada, New Zealand, and South Africa). This group is characterized by the existence of a small number of banks with branches throughout the country, holding a significant part of total deposits, along with a relatively large number of smaller banks that are regional or local in emphasis. Such systems exist in France, Germany, and India. Japan has a small number of large city banks with branch networks but a larger number of local banks.

France.

Banking institutions in France were classified after World War II into three main groups: deposit banks, banques d'affaires (or investment banks), and institutions that were either specialized or operated mainly outside France. New banking legislation in 1966 greatly reduced the importance of the distinction between deposit banks and banques d'affaires. There was also (1) a further concentration of banking resources, as a result of several large mergers and also of greater financial integration through share-exchange agreements and interlocking directorates, and (2) the conversion of a number of banques d'affaires into deposit banks, which hived off their investment interests into separate investment or holding companies.

Further legislation in 1982 nationalized the remaining large and medium-sized banks (36 in all, plus two financial holding companies - those of Indosuez and Paribas); the largest deposit banks had already been nationalized after World War II. Another new law in 1984 abolished the old divisions between the several categories of banks, which were now defined simply as établissements de crédit, able to receive deposits from the public, undertake credit operations (including loans), and provide means of payment. The intention was to move cautiously

toward a system of "universal banking." The new law was extended to cover the Caisse Nationale de Crédit Agricole, the banques populaires, the crédit mutuel, the central organizations of the cooperatives and the savings banks (and thereby institutions affiliated with them), and semipublic institutions like the Crédit Foncier and the Crédit National, but not the Caisse des Dépôts et Consignations nor the central banking institutions.

All the regional banks and some local banks have branches. The balanced character of the regional economies often provides these banks with a good portfolio of risks; they serve not only a prosperous agriculture but also a number of local industries. Some of the local banks are also very sound institutions, despite their small size.

The survival of a hybrid system in France, despite the long-run trend toward centralization, reflects certain characteristics of French society. These included, until recently, a strong emphasis on small business, together with a preference for individual and personal service. Particularism in some parts of France manifests itself in support for local institutions, and the local banker also often has the advantage of special knowledge of local industries and people, which makes possible the acceptance of risks that the big banks decline.

Germany.

An even more direct conflict between the forces favouring concentration and those working against it may be seen in Germany, where banking grew in the latter part of the 19th century along with industry. The banks were inclined to rely mainly on their own capital resources and did not at first try to attract deposits from the public. Not until 1874 did the Deutsche Bank A.G. begin to seek deposits through offices specially opened for the purpose. This was done to provide cheap finance for traders, the deposits being invested in mercantile bills that were regarded as both safe and liquid. In pursuit of deposits, the banks built up a widespread network of branch offices, which were also used to establish and maintain industrial contacts throughout the country. The unification of Germany in 1871 removed the political obstacles to a more integrated banking system, and the

selection of Berlin as the capital made that city the country's financial centre. Four of the largest banks were already established there; the new Reichsbank was set up in 1876. In addition, the larger and more enterprising of the provincial banks were attracted to the capital. The Berlin stock exchange rapidly displaced that of Frankfurt am Main as the country's leading securities market.

The Berlin banks extended their influence by developing correspondent relationships and subsequently by acquiring a financial interest in the provincial banks and being represented on their boards. Each of the big Berlin banks came to be associated with a group of provincial banks more or less under its control. At the same time, all of the banks, Berlin and provincial alike, expanded their business by opening branches.

During World War I the degree of centralization increased; by 1918 the big Berlin banks held more than 65 percent of total deposits. In the early 1920s there were amalgamations, and branch systems became much larger. Bank failures and the financial crisis of 1931 resulted in further consolidation until the German banking system was dominated by three giants. But there were countervailing forces. Probably the most important of these was the establishment of publicly owned banking institutions, such as the communal savings banks and their central institutions, the Girozentralen, which became of increasing importance after World War II.

German savings banks, which were permitted to have checks drawn on them from 1909 and which had giro clearing from the 1920s, now offer a wide range of services, especially to lower income groups and smaller businesses. The large commercial banks have concerned themselves more with big business and with wealthy individuals. The savings banks now compete in wholesale banking as well. A number of them, together with their Girozentralen, are to all intents and purposes "universal banks," like the Big Three and the larger regional banks. The Big Three (the Deutsche Bank, the Dresdner Bank, and the Commerzbank) remain unchallenged only in stock exchange and foreign banking business.

Of the private bankers, only about a half-dozen are of any size. The bigger private banks are important in the fields of investment and wholesale banking, while the smaller ones flourish in the leading stock-exchange cities, such as Düsseldorf and Frankfurt am Main. Many of these private bankers, however, are not bankers in the true sense; they subsist mainly on stock-exchange transactions, investment services, portfolio management, and insurance and mortgage brokerage. There are also consumer finance institutions, mortgage and other specialist banks, and a large number of cooperatives.

Regional and private banks are often within the sphere of influence of the Big Three. In some cases the latter have a financial interest in these banks, and in some cases they own them. The Big Three also have shares in certain of the private mortgage banks. There are also "cooperation agreements," and a number of mergers have taken place. In these several ways, much more integration exists than appears on the surface. While banking in Germany remains a hybrid system, a trend toward greater concentration is evident.

India.

Until the 1950s, banking in India was carried on by a large number of banks, many of them quite small. India is still primarily an agricultural country, with an economic and social structure based largely on the village. The integration of banking has been impeded by poor communications, by illiteracy, and by the barriers of language and caste. Banking and credit have remained largely in the hands of the so-called indigenous banker and the village moneylender. Although their influence has been greatly reduced in recent years, they still remain important in many an up-country area. The indigenous banker, who is also a merchant, offers genuine banking services: accepting deposits and remitting funds; making loans quickly and with a minimum of formality; and, by means of the hundi (a credit instrument in the form of a bill of exchange), financing a still significant, if declining, portion of India's internal trade and commerce.

Efforts were made to eliminate the moneylender by developing a network of rural credit cooperatives. When progress proved to be slow, a more successful

alternative was found in requiring banks to open "pioneer" branches in rural areas. The first branches were those of the semipublic Imperial Bank of India and its nationalized successor, the State Bank of India (and its subsidiaries). Many smaller banks began to disappear, sometimes by merger and sometimes as a result of failure. Between 1952 and 1967 the number of "reporting" banks fell from 517 to 90. Nationalized banks, including the State Bank of India and its seven subsidiaries, the 14 large commercial banks taken over in 1969, and the six additional banks nationalized in 1980, accounted for more than 90 percent of aggregate deposits in commercial banks. Banking services are also provided by chit funds, which accept and pay interest on monthly deposits against which it is possible to draw only by way of loan, and by Nidhis, mutual loan societies that have developed into semibanking institutions but deal only with their member shareholders.

The main path of banking development in India is the expansion of bank branches into the under-banked areas. The authorities have sought to expand the number of branches but to avoid their concentration in the larger towns and cities and, in particular, to provide the rural areas with adequate facilities. The ultimate objective is to encourage the mobilization of deposits on a massive scale throughout the country, a formidable challenge in a country of 575,000 villages, and a stepping up of lending to weak sectors of the economy.

Japan.

Banking business in Japan is largely concentrated in the hands of the big banks (some of which are specialized), though a number of small banks still survive. The principal classes of banks are city banks and regional banks, but it should be noted that the distinction has no legal basis, though they are separately supervised. Both belong to the Federation of Bankers' Associations of Japan. The city banks service mainly manufacturing industry and commerce, particularly the big firms, while the regional banks are based on a prefecture, though some extend their operations into neighbouring prefectures, collecting deposits and lending to local businesses and smaller firms. The regional banks have city bank correspondents, not only to hold

surplus balances but also for assistance in investing their funds, especially in the call-money market. In addition, a city bank may introduce certain of its large customers to a regional bank (e.g., a big company having a local factory). City correspondents in Japan do not, however, provide the wide range of ancillary services common in the United States.

Since World War II there has been much stability in Japanese banking, but the city banks have suffered a relative decline in the importance of their business in competition with other institutions, especially the agricultural cooperatives, which attract the larger part of the Treasury's payments owing to government purchases of the rice crop. There has also been a relative increase in the importance of the life insurance companies and the trust funds, which have attracted sizable funds from the general public.

BANKING IN PLANNED ECONOMIES

The old Soviet banking system was established by the credit reforms of the early 1930s, which centralized practically all short-term credit in the hands of the Gosbank (State Bank, established in 1921). There was much restructuring of banking during succeeding years, mainly to ensure that the system became an effective instrument for carrying out the national economic plan. The Gosbank's control over payments flows was also tightened in order to maintain stability of prices. The activities of the Gosbank are by no means limited to purely financial operations, and it actively controls the implementation of production and financial plans and the spending of wages funds. In addition, it has a monopoly of the note issue and is responsible for putting money into circulation.

The Gosbank was originally concerned with the provision of short-term credit; a number of other banks were created to finance capital investment in the socialized economy. Even in the 1930s there was a tendency to consolidate these banking units, and this continued into the postwar period. In 1959 there were further mergers and a reallocation of activity, as a result of which the Stroybank emerged as the credit and financial institution responsible for canalizing state budgetary appropriations and also short- and long-term credits into capital investment in

various sectors of the economy. The savings bank system, with 79,000 branches, became part of the Gosbank in 1963. The only other banking institution was the Vneshekonombank (Bank for Foreign Trade), whose operations were considerably expanded in 1961. Originally concerned mainly with providing currency for tourists and diplomatic missions and with remittances from abroad, the Vneshekonombank came to handle all foreign-exchange transactions, including those relating to trade.

In 1988 the credit banking system was again broken up into smaller units: Vsesoyuzny Bank Dlya Finansirovaniya Promyshylennosti i Kapitalnykh Vlozheny (All-Union Bank for the Financing of Capital Investments), Agroprombank (Agro-Industrial Bank), and Zhilsotsbank (Bank for Housing, Municipal Services, and Social Development). Also in 1988, Sberegatelno-kreditny Bank (Saving and Consumer Credit Bank) assumed responsibility for public and consumer deposits and loans.

The Gosbank operated as follows: it had a policymaking head office and principal offices in the various Soviet republics. There were also regional offices and a network of 4,500 local branches; the latter were the bank's main points of contact with a variety of economic enterprises, collective farms, and lower-level government units. The Gosbank served the needs of the urban population through its network of branches, which collected rent, taxes, and other compulsory payments and contributions. It maintained a small number of special cash service agencies in large industrial establishments and at construction projects. Seasonal agencies operated at remote places where large purchases of farm products were made.

In the industrial area, the Gosbank served hundreds of thousands of state enterprises that operated on the basis of cost accounting. Each of these enterprises had its own working capital and prepared a balance sheet and a statement of income; it was permitted to borrow regularly or occasionally, depending on its needs. The Gosbank had hundreds of thousands of other customers comprising collective farms, party, trade union, cultural and other organizations, and

individuals. The aggregate balances maintained by Gosbank customers were small in comparison with the cash balances held by business and government accounts in the United States.

Transactions of individuals.

Rising incomes, an increase in savings, and the provision of facilities for crediting wages to savings accounts and making periodic payments from them resulted in a rapid increase in the transactions of individuals. Housing loans and tourism had also grown in importance. Savings bank offices handled virtually all the accounts and transactions of individuals.

Transactions of collective farms.

An even greater problem was created by the rise in transactions involving the collective farms. Before 1953 the collectives paid in kind for the services performed by the state-owned tractor and farm-machinery stations. After 1953 these transactions gradually shifted to payment in cash. As a first step, the tractor and machinery stations were closed down and their equipment sold to the collectives. The farms then had to pay in cash for all machinery and fuel, as well as for building materials, fertilizers, and other supplies. Subsequently, they were enabled to sell their output to the state for money. Farm labour likewise came to be remunerated in cash rather than in kind. In 1953 only one-third of the "compensation" for work contributed by members of the collectives was in cash; by 1963 the proportion was nearly three-quarters. In 1965 the flow of money income to the farm population was increased further by the introduction of state pensions for collective farmers. In 1966 it was decided to make minimum monthly payments to the members of collective farms, and this resulted in an even greater use of money in the farm sector.

The growth of money flows and of bank lending in rural areas, as well as the need to service a growing clientele in the villages, greatly added to the complexity of Gosbank operations, which had been geared primarily to the needs of industry and government. The Gosbank attempted to resolve its problems by simplifying

payments procedures, through its efforts to work out a system of offsetting mutual claims were not initially very successful.

After the credit reforms of 1930-32, a uniform system of interest rates was applied to all short-term credits, irrespective of the purpose of the loan or the financial condition of the borrower. Higher rates were charged as a penalty on overdue loans. In the 1960s there was a move to differentiate rates, it being accepted as a matter of principle that bank funds should be more expensive than an enterprise's own working capital and that borrowings made necessary by shortcomings of management (for example, excessive inventories or the erosion of working capital) should carry higher rates than loans to finance normal needs. Penalty rates for overdue loans and for late payments were also increased, and collection was more vigorously enforced. Until the country's dissolution in 1991-92, interest rates in the Soviet economy were an integral part of the relationship between the bank and the enterprises that it both served and regulated.

Economic Growth and Planning

Introduction

Economic growth involves increases over time in the volume of a country's per capita gross national product (GNP) of goods and services. Such continuing increases can raise average living standards substantially and provide a stronger base for other policy objectives such as national defense, various kinds of capital investments, or public welfare services. It is only in the last two centuries that continued growth in living standards has been realized for a number of now-developed countries, and this process has broadened in the 20th century to include a number of developing countries. However, the fairly steady expansion in the third quarter of the 20th century gave way to a period of slower and more erratic growth for both high- and low-income countries, while some of the economically poorest countries were thus far unable to establish a self-sustaining pattern of development. It also became increasingly evident that there were serious environmental problems associated with some types of growth in production.

This article examines the record of economic growth and development, some explanations for the changes involved, and the attempts by governments to plan these changes. Five major issues are involved. The first is why economic growth occurs more quickly in some countries and periods than in others. It is the increase in the size and quality of the factors of production that underlies growth, but certain forces - innovations and entrepreneurship, the part played by governments, and the role of investment as distinct from consumption - deserve special attention. A variety of models of economic growth give expression to the understanding of these forces. Increasing attention has been paid in these models and in policy to the international aspects of growth. This trend is partly a reflection of the growing internationalization of economic activity. It also reflects a number of potentially destabilizing changes in the international economy that became evident during the 1970s. While the precise nature of their effects is open to debate, among these changes should be noted the transition to more flexible exchange rates, the supply

shocks in petroleum and other products, the growth of international debt, and the development of several major centres of economic power.

A second issue is the challenges facing the low-income countries, namely, to move from subsistence levels of per capita income to a level that would generate self-sustaining growth and also to reduce the gap between themselves and the higher-income countries. Differences among the lower-income countries warn against making sweeping generalizations on the development process, but three topics have attracted much attention. One is how far existing private and public organizations must be changed so as to institutionalize development. A second is the view that, particularly for manufactures and for smaller markets, reliance on import substitution should give way fairly quickly to export development. The third is the impact of population growth on both development and living standards.

The uneven patterns of growth and development have led to many strains. In the case of higher-income countries these have appeared particularly as declines and failures in some of their older industries in the face of increased competition among themselves and with the newly industrializing countries. In many developing countries there have been repeated calls for a new organization of world institutions geared to a more equitable distribution of wealth.

A third issue, productivity, is central to changes in living standards and to the analysis of international competitiveness. Productivity is the ratio of what is produced to what is required to produce it, a ratio beset with measurement problems. Studies of the reasons for the growth of productivity have focused on technological change and on the accumulation of physical and human capital. However, a considerable number of social, political, and economic issues appear to be involved in the striking differences in rates of change in productivity among countries. Nor is there agreement on the reasons for and significance of the marked slowing of productivity growth in many countries during the 1970s and 1980s. Some economists see merely passing factors at work, while others predict continuing problems for developed countries and the international economy generally.

A fourth major issue is the attempt to maintain growth and increase development through economic planning. Such planning has other objectives as well, such as regional development, constraining wage-price inflation, and easing the adjustment between declining and growing sectors. Planning became a widespread phenomenon during and just after World War II and was given further emphasis in many newly independent countries that were industrializing. The degree of detail and the methods used in planning range from heavily centralized direction and substantial public ownership to attempts simply to trim the overall balance of supply and demand, with different degrees of attention to industrial strategies along the way. Beginning in the 1970s the emphasis shifted to more decentralized planning, with deregulation and privatization of industry as two aspects of this process.

Underlying economic growth and planning is a fifth issue, the attempt to predict economic activity. Modern forecasting involves a variety of computer-based techniques at the level of the firm, the country, and the international economy. The accuracy of forecasting has been reduced by increased uncertainty in the global and national economies since the early 1970s.

How economies grow.

Growth can best be described as a process of transformation. Whether one examines an economy that is already modern and industrialized or an economy at an earlier stage of development, one finds that the process of growth is uneven and unbalanced. Economic historians have attempted to develop a theory of stages through which each economy must pass as it grows. Early writers, given to metaphor, often stressed the resemblance between the evolutionary character of economic development and human life - e.g., growth, maturity, and decadence. Later writers, such as the Australian economist Colin Clark, have stressed the dominance of different sectors of an economy at different stages of its development and modernization. For Clark, development is a process of successive domination by primary (agriculture), secondary (manufacturing), and tertiary (trade and service) production. For the American economist W.W. Rostow, growth

proceeds from a traditional society to a transitional one (in which the foundations for growth are developed), to the "take-off" society (in which development accelerates), to the mature society. Various theories have been advanced to explain the movement from one stage to the next. Entrepreneurship and investment are the two factors most often singled out as critical.

Economic growth is usually distinguished from economic development, the latter term being restricted to economies that are close to the subsistence level. The term economic growth is applied to economies already experiencing rising per capita incomes. In Rostow's phraseology economic growth begins somewhere between the stage of take-off and the stage of maturity; or in Clark's terms, between the stage dominated by primary and the stage dominated by secondary production. The most striking aspect in such development is generally the enormous decrease in the proportion of the labour force employed in agriculture. There are other aspects of growth. The decline in agriculture and the rise of industry and services has led to concentration of the population in cities, first in what has come to be described as the "core city" and later in the suburbs. In earlier years public utility investment (including investment in transportation) was more important than manufacturing investment, but in the course of growth this relationship was reversed. There has also been a rise in the importance of durable consumer goods in total output. In the U.S. experience, the rate of growth of capital goods production at first exceeded the rate of growth of total output, but later this too was reversed. Likewise, business construction or plant expenditures loomed large in the earlier period as an object of business investment compared to the recent era. Whether other countries will go through the same experience at similar stages in their growth remains to be seen.

Comparative growth rates for a group of developed countries show how uneven the process of growth can be. Partly this unevenness reflects the extraordinary nature of the 1913-50 period, which included two major wars and a severe and prolonged depression. There are sizable differences, however, in the growth rates of the various countries as between the 1870-1913 and 1950-73 periods and the period

since 1973. For the most part, these differences indicate an acceleration in rates of growth from the first to the second period and a marked slowdown in growth rates from the second to the current period. Many writers have attributed this to the more rapid growth of business investment during the middle of the three periods.

The relatively high rates of growth for West Germany, Japan, and Italy in the post-World War II period have stimulated a good deal of discussion. It is often argued that "late starters" can grow faster because they can borrow advanced technology from the early starters. In this way they leapfrog some of the stages of development that the early starters were forced to move through. This argument is nothing more than the assertion that late starters will grow rapidly during the period when they are modernizing. Italy did not succeed in growing rapidly and thereby modernizing until after World War II. Together with Japan and Germany it also experienced a large amount of war damage. This has an effect similar to starting late, since recovery from war entails building a stock of capital that will, other things being equal, embody the most advanced technology and therefore be more productive and allow faster growth. The other part of this argument is the assertion that early starters are actually deterred from introducing on a broad front the new technology they themselves have developed. For example, firms in a country that industrialized early may be inhibited from introducing a more modern and efficient means of transportation on a broad scale because there is no guarantee that other firms handling the ancillary loading and unloading tasks will also modernize to make the change profitable.

Related to this is the problem of whether or not per capita income levels and their rates of growth in developed economies will eventually converge or diverge. For example, as per capita incomes of fast growers like the Italians and Japanese approach those of economies that developed earlier, such as the American and British, will the growth rates of the former slow down? Economists who answer in the affirmative stress the similarities in the changing patterns of demand as per capita income rises. This emphasis in turn implies that there is less and less chance to borrow technology from the industrial leaders as the income levels of the late

starters approach those of the more affluent. Moreover, rising per capita incomes in an affluent society usually are accompanied by a shift in demand toward services. Therefore, so this argument goes, differences in income levels and growth rates between countries should eventually narrow because of the low growth in productivity in the service sector. The evidence is inconclusive. On the one hand, growth is a function of something more than the ability to borrow the latest technology; on the other hand, it is not clear that productivity must always grow at a slower rate in the service industries.

A rapidly increasing population is not clearly either an advantage or a disadvantage to economic growth. The American Simon Kuznets and other investigators have found little association between rates of population growth and rates of growth of GNP per capita. Some of the fastest growing economies have been those with stable populations. And in the United States, where the rate of growth of population has shown a downward historical trend, the rate of growth of GNP per capita has increased over the last century and a half. Another finding by Kuznets is that while GNP per capita in 1960 was substantially higher in the United States than in any European country, there was no significant difference in the per capita growth rates of all these countries over the period 1840 to 1960 as a whole. The conclusion is that the United States started from a higher per capita base; this may have been the result of its superior natural resources, especially its fertile agricultural land.

THE ANALYSIS OF GROWTH

To explain why some countries grow more rapidly than others or why a country may grow more rapidly during one period of history than another, economists have found it convenient to think in terms of a “production function." This is a mathematical way of relating some measure of output, such as GNP, to the inputs required to produce it. For example, it is possible to relate GNP to the size of the labour force measured in man-hours, to capital stock measured in dollars, and to various other inputs that are considered important. An equation can be written that states that the rate of growth of GNP depends upon the rates of growth of the

labour force, the capital stock, and other variables. A common procedure is to assume that the influence of the separate inputs is additive - i.e., that the increase in the growth of output caused by increasing the rate of growth of, say, capital is independent of the rate of growth of the labour force. This is the starting point of a great deal of current empirical work that attempts to quantify the importance of different inputs.

Under certain assumptions, some reasonable and some patently false, it is possible to conclude that what labour and capital receive in the form of wages, profits, and interest is a fair measure of what they contribute to the productive process. Thus in the United States in the period following World War II the share of output going to labour was approximately 79 percent, while the share of output distributed as "profits" was 21 percent. If we assume that these proportions determine how much we should weight the rate of growth of the labour force and of capital respectively in determining their contribution to the rate of growth of output, we must conclude that the relative contribution of capital is slight. Alternatively, we may say that some given percentage increase in the rate of growth of the labour force will have a much larger influence on the rate of growth of output than the same percentage increase in the rate of growth of capital. This is a puzzling result and can be traced to the assumption that the influence of separate inputs is additive.

Quality improvements in the inputs.

Much work has been done in an effort to measure the inputs in the productive process more accurately by taking account of improvements in the quality of both labour and capital over time. For example, it has been argued that the amount of a worker's time spent on his formal education is positively related to the income he receives and to his productive contribution. Measuring the number of man-hours worked from one period to the next will not give a true picture of the increase in labour input if the average amount of education received by workers is changing. Man-hour units must be converted to "efficiency" units. Thus if a labour force of 100 workers in the first year all had an eighth-grade education, while 20 years later each member had a 10th-grade education, then measured in efficiency units the

labour force had grown. If the length of time spent on formal education increases over time, then the growth of the labour input will be larger if measured in efficiency units. There is, thus, an element of capital in the labour force.

Examples of investment in human capital are expenditures on health and on all types of education, including on-the-job training. Expenditures of this sort increase the quality of the labour force and its ability to perform productive tasks. Many economists have argued that technological progress is really nothing but quality improvements in human beings. Some economists take an even broader view and speak of the "production of knowledge" as the clue to technological progress. The production of knowledge is a broad category including outlays on all forms of education, on basic research, and on the more applied type of research associated especially with industry. It is argued that fast-growing industries tend to be those having a high research and development component in their total costs. In addition, firms within an industry that have large research and development budgets tend to experience the most rapid technological progress. The argument is that technical change and improvements must originate in inventions that lead to innovations in the products produced or in the processes whereby existing products are manufactured.

A similar argument applies to the size of the capital stock. It can be maintained that design improvements increase the efficiency of capital goods so that a dollar's worth of machinery purchased today may be much more efficient than a dollar's worth of depreciated machinery purchased yesterday. The rate of growth of the capital stock measured so as to take account of quality improvements will be greater than the rate of growth of the capital stock measured in a way that neglects the differences between "vintages."

Some economists have stressed "economies of scale." For example, if an increase in the use of capital and labour leads to a greater than proportionate increase in output, this is said to result from economies of scale. Economies of scale may arise because an expansion of the market justifies a radical change in productive techniques. These new techniques may be so much more efficient that the returns

in the way of increased output are much greater proportionately than the increase in inputs.

Another source of growth and of technical progress in particular has been seen in shifts of demand from low productivity sectors to high productivity sectors, thus causing resources to be reallocated. The most notable movement has been the shift of resources, especially labour, out of agriculture - a traditionally low-productivity sector. Such shifts act to increase the rate of growth of output in ways that cannot be accounted for by simply measuring growth in total inputs. Historically, the allocation of both capital and labour have shifted during the growth process from low productivity sectors to high ones, causing the rate of growth of output to exceed the weighted average of the rates of growth of total inputs.

Entrepreneurship.

This historical fact points to an element that has received little attention so far: the influence of entrepreneurship. If the allocation of resources changes during the course of growth and development, it does so under the leadership of an entrepreneurial class. The quality of entrepreneurship is seen by many economists as an important explanation of differences in the rate of technical progress between countries. Decisions must be made somewhere along the line as to whether a new product or process will be introduced. It has been argued that two countries undertaking similar amounts of investment leading to more or less identical rates of growth in the capital stock will not necessarily show the same rate of technical progress. In one country entrepreneurs may be undertaking enterprise investment that has as its aim the introduction of the most advanced types of production techniques, those that will lead to a rapid growth of labour productivity. In the other, because of hesitation or ignorance, the investment program may lead only to marginal changes in productive processes; the resulting growth in labour productivity and GNP will be small. For example, much has been said since World War II about the more aggressive nature of German businessmen as compared to their English counterparts. The emphasis on the role of the entrepreneur in

economic growth stems from the theoretical work of the economist Joseph A. Schumpeter, but many others have echoed it.

The play of influences.

Much thinking assumes, then, that contributions to output from growth of individual inputs are independent of one another. This assumption allows many growth theorists to conclude that capital investment is relatively unimportant as a growth factor. If there is interaction between the rates of growth of the different inputs, however, then it is possible to draw different conclusions. For example, over time there are likely to be improvements in the quality of capital goods. A machine that requires so much steel and so much labour to manufacture may be twice as productive as an older machine that required the same amount of raw materials and labour in its manufacture. Thus the rate of growth of technical progress and the rate of growth of the capital stock measured in natural units interact. Furthermore, the interaction between technical progress and capital formation is not necessarily in one direction. New knowledge opens up new production possibilities and gives rise to potential increases in technical progress and profits. Or the better educated the labour force, the more adaptable it is likely to be and therefore the better able to cope with new production techniques. At the same time, the higher the rate of growth of capital, the higher will be the growth of incomes and therefore the demand for education. The fact that much of the overall growth of technical progress stems from the transfer of resources and the positive association between the rate of transfer of resources and the rate of growth of the capital stock is another example of interdependence or complementarity between the growth of the inputs. But, again, capital investment undertaken to develop new lines of production will also be dependent on technological progress going on in those areas.

Conventional marginal productivity doctrine argues that as an input such as capital rises relative to labour, the additional output or marginal product that can be attributed to this extra amount of capital will be less than what a unit of capital on the average had been producing before. Marginal productivity doctrine also

assumes that each unit of capital is identical with the next. This assumption is the basis for the argument that as more units of capital are utilized in production with a given amount of labour, it will push down the former's marginal product. There is the possibility, however, that additional units of capital may enhance the productivity of existing units: for example, an increase in the amount of capital resources devoted to the development of transportation and distribution may raise the productivity of capital employed, say, in manufacturing. The development of this kind of social overhead capital is certainly a prerequisite for a high return to capital in manufacturing, wholesaling, and retailing.

The analysis can be carried back one more step, to the basic determinants of growth. Economists ask why it is that capital, labour, or technical progress has grown more rapidly in one economy than in another or at one time than at another. Historically, the transition from a subsistence-level, underdeveloped state to a higher-level, developed one has been accompanied by a decline in the death rate followed by a decline in the birth rate. This has the effect of first speeding up the rate of growth of the population and labour force and then reducing it as birth rates fall. Migration can alter this picture, often unpredictably. In the United States, for example, the rate of growth of the population and labour force during the 19th and early 20th centuries was higher than in most other developed countries, mainly because of high rates of immigration. From 1840 to 1930, the native-born U.S. population increased about 600 percent, while the number of those of foreign birth increased 1,300 percent.

The role of government.

The differences in rates of growth are often attributed to two factors: government and entrepreneurship. The two are not mutually exclusive. In the early stages of sustained growth, government has often provided the incentives for entrepreneurship to take hold. In some economies the development of transportation, power, and other utilities has been carried out by the government. In others the government has offered financial inducements and subsidies. The land given U.S. railroad developers in the second half of the 19th century is a notable

example of the latter. Another important role governments have played in the early stages is to help establish the sort of capital and money markets in which lenders could have confidence. Without financial intermediaries acting as brokers between lenders and business borrowers, it is difficult to envisage economic growth taking place on a sustained and rapid basis.

In the 19th century most liberal thinkers held that the main role for government in a developed capitalist system was that of a policeman: to preserve law and order, uphold the sanctity of private property, and give business as much freedom as possible. The Great Depression of the 1930s persuaded many that a laissez-faire system did not automatically provide the necessary incentives to the innovation and risk bearing essential for economic growth. This led to a good deal of writing on the role that governments might play in stimulating growth. Economists have argued that, at the very least, governments can undertake to prevent serious and prolonged recessions. Only in this way can a general business psychology be developed that assumes growth to be the natural course of things, so that investment programs will pay off.

Growth theorists since World War II have gone further, arguing that it is not enough simply to achieve full employment periodically. Some maintain that it is necessary to maintain full employment over an extended period of time if high growth is to result. This argument relates to the earlier point that two economies may experience the same rate of growth of capital but that overall growth and technical progress will proceed at a much more rapid rate in one than in the other because of differences in the quality of new capital goods produced. The term enterprise investment has been used to describe the kind of capital formation that involves innovations and that by building ahead of demand generates rapid rates of growth of productivity or technical progress. But to get such growth, it has been argued, an economy must be run "flat out," at full speed. While this has been subject to some dispute, there is a fairly general consensus that growth will be faster when unemployment fluctuates within a narrow range and at low levels.

A variation on this argument is the question of how a government may intervene to determine the distribution of output between those types of expenditure that contribute to growth and those that lead to the immediate satisfaction of consumer demand. Here the choice lies between business investment, research, and education on the one hand and consumption on the other. The larger the first three, the more rapid will be the rate of growth. Governments giving a high priority to growth have various means at their disposal for influencing it. Consumption can and has been constrained through increases in income tax rates. The same is true of other tax rates such as the property tax - the chief revenue source for primary and secondary education in the United States. Tax credit for research and development expenditures is a common method for encouraging business outlays that may lead to innovations. The same method has been used to stimulate business investment outlays. "Easy money" policies on the part of the central bank, whereby the cost of borrowed funds and their availability are indirectly regulated in such a way as to encourage business borrowing, may lead to higher levels of real investment.

The true cost of stimulating growth will always be a temporary cut in current consumption. Only in the future can the economic benefits of the higher investment be realized. By the same token, current consumption can always be enlarged by a neglect of the future. It is even possible for current production to be so biased toward the satisfaction of immediate needs that the productive capacity of an economy slowly declines as capital goods are not replaced. Between the extremes of total neglect of future generations and the paring down of current consumption to a bare subsistence minimum lie an infinite number of possibilities.

THE SOCIAL COST OF GROWTH

The belief that governments should have a large say in choosing the "right" rate of growth has also led some writers to challenge the social and economic value of economic growth in an advanced industrial society. They attribute to growth such undesirable side effects of industrialization as traffic congestion, the increasing pollution of air and water, the despoiling of the landscape, and a general decline in man's ability to enjoy the "real" amenities of life. As has been seen, growth is

really a transformation whereby certain industries experience a rise in importance followed by an eventual decline as the market for their output becomes relatively saturated. Demand, relatively speaking, moves on to other types of industries and products. All of this naturally implies a reallocation of resources over time. The faster these resources move, other things being equal, the more rapidly can growth and transformation proceed. The argument can be recast in terms of this transformation. A slower rate of growth in per capita consumption will slow down the rate of transfer of resources, but it may also result in a more livable environment. The rate of growth of individual welfare, so measured as to take into account non-consumable amenities, may even be increased. Some argue that in a growth-oriented society wants are created faster than the industrial machine can satisfy them, so that people are more dissatisfied and insecure than they would be if growth were not given such a high value. It is held by some critics that, in modern industrial society, consumption exists for the sake of justifying production rather than production being carried out to satisfy consumer desires. These arguments are a powerful challenge to those who see growth as the most important economic goal of a modern society.

THEORIES OF GROWTH

In discussing theories of growth a distinction must be made between theories designed to explain growth (or the lack of growth) in countries that are already developed and those concerned with countries trapped in circumstances of poverty. Most of what follows will be confined to the former.

As the British economist John Maynard Keynes pointed out in the 1930s, saving and investment are not usually done by the same persons. The desire to save does not necessarily generate investment. If savers attempt to save a larger share of their income than before (thereby consuming less) and if this is not matched by an equal increase in the desire of others to invest, total spending will decline. A natural reaction on the part of business will be to cut back on production, thereby reducing incomes earned in production. The final effect may be a cumulative movement downward as total demand becomes insufficient to employ all of the labour force.

This break in the circular flow of income and expenditure suggests the possibility of a capitalist economy alternately experiencing periods of prolonged and severe unemployment (when desired savings at full employment exceed what the economy wishes to invest at full employment) and periods of serious inflation (when the inequality is reversed). This situation had not been the case historically for developed economies until the early 1970s. In the following discussion, some attention will be paid to the ways in which the various theories of growth account for this important historical fact.

Role of the entrepreneur.

Modern growth theory can be said to have started with Joseph A. Schumpeter. Unlike most Keynesian or pre-Keynesian theorists, Schumpeter laid primary stress on the role of the entrepreneur, or businessman. It was the quality of his performance that determined whether capital would grow rapidly or slowly and whether this growth would involve innovation and change - i.e., the development of new products and new productive techniques. Differences in growth rates between countries and between different periods in any one country could be traced largely to the quality of entrepreneurship. The latter in turn reflected certain historical and cultural values carried by the business class. Schumpeter also attributed much of the growth of technical progress and of the supply of labour to the entrepreneur. Thus, in more modern terminology, Schumpeter's explanation of why demand and supply have grown more or less at the same rate would be that supply adjusted to demand while demand in turn reflected the activities and investments of the entrepreneur.

Schumpeter believed that capitalism by its very success "sows the seeds of its own destruction." The American economist Alvin H. Hansen argued in the late 1930s that capitalism was in trouble in the United States for other reasons. According to Hansen, the closing of the geographic frontier, the decline in the rate of population growth, and the capital-saving character of recent innovations had all worked to increase the likelihood of stagnation by reducing the need for investment. The savings available in a mature economy would tend to exceed the amount that the

economy would want to invest (at levels of full employment) and by progressively larger amounts as time went on. This condition naturally would lead to increasing rates of unemployment as the discrepancy between demand and potential output widened. Hansen's views were very much coloured by the economic conditions of the 1930s. The record of the three decades after World War II did much to overcome the pessimism generated by the Great Depression.

The role of investment.

In Keynes's General Theory, investment played a key role in that it was presented as the most important factor governing the level of spending in an economy, despite the fact that it typically was only one-fifth to one-sixth of total spending. This paradox can be understood in terms of a concept also developed in the 1930s, the multiplier. The multiplier was the amount by which a change in investment would be multiplied in achieving its final effect on incomes or expenditures. If, for example, investment increases by $10, the extra $10 of expenditures will generate, assuming unemployed resources, an extra $10 of production and subsequently incomes in the form of wages and profit. This increase, however, is hardly the end of the matter since most of the additional incomes earned will be respent on consumer goods. If nine-tenths of any change in income is spent on consumer goods and one-tenth is saved, consumption will increase by $9. But again, one person's expenditures are another person's income, so that incomes now rise by $9 of which $8.1 is respent on consumer goods. The process continues until expenditures, incomes, and production have increased by $100, of which $90 is consumption and $10 the original change in investment. In this case the multiplier is 10.

But investment may be a source of instability if it is not maintained at a rate sufficient to stimulate demand for the production it is creating. Is there any guarantee that supply or productive capacity will grow at the same rate as demand so that neither excess capacity nor excess demand results? The British economist R.F. Harrod and the American economist E.D. Domar put this question in a very simple mathematical form. In their equations, the rate of growth of supply (i.e., the

production function as defined above) is equal to the rate of growth of capital stock. Through investment this capital stock is augmented. The rate of growth of demand depends upon the rate of growth of investment or, more correctly, upon the rate of growth of nonconsumption expenditures. Thus investment affects both demand and supply. But the Harrod-Domar analysis still did not answer the question of what kept the system from becoming increasingly unstable.

Demand and supply.

Much contemporary growth theory can be viewed as an attempt to develop a theoretical model that would bring the rate of growth of demand and the rate of growth of supply into line, since a model implying that capitalist systems are inherently unstable would not correspond to the historical facts. Models of growth may be classified according to whether they emphasize adjustments in demand (supply-determined models) or adjustments in supply (demand-determined models). One of the better-known examples of the supply-determined model was developed by the British economist J.R. Hicks. Hicks assumed that the spending propensities of consumers and investors were such as to cause demand to grow at a rate in excess of the rate of growth of maximum output. This assumption meant that during any "boom" the economy would eventually run into a "ceiling" that, while also moving upward, was moving less rapidly than demand. The long-run rate of growth of the economy would be determined by the rate of ascent of the ceiling, which in turn would depend upon supply factors such as the rate of growth of the labour force and the rate of growth of technical progress or productivity. If for some reason these were to grow more rapidly, then output would also grow more rapidly as demand adjusted upward to the more rapid growth of supply.

An example of a demand-determined model of growth is one developed by the American economist J.S. Duesenberry. In the Duesenberry model, spending propensities of consumers and investors are such as to generate steady growth in demand. Assume that instead of spending nine-tenths of any change in income on consumer goods, as in the multiplier example above, they choose to spend 0.95. This increase will cause the rate of growth of demand to increase. The question is

whether it will also cause the rate of growth of production to increase or whether it will merely result in price increases. If productivity or technical progress responds to a higher rate of growth of demand, as Duesenberry assumes, then production can grow more rapidly. Although in both the Hicks and Duesenberry models demand and supply grow at the same rate, the adjustment mechanisms are entirely different. In the Duesenberry model supply adjusts to demand; in the Hicks model demand adjusts to supply.

Other models of growth also illustrate this distinction between demand-determined and supply-determined growth. The British economist N. Kaldor assumed that there is a mechanism at work generating full employment. Simply stated, in his model an inadequate rate of investment will be offset by shifts in the distribution of income between profits and wages, which will cause consumption to change in a compensating manner so that overall demand is unchanged. While there are important differences between the Hicks and Kaldor models, both can be described as models of supply-determined growth.

Another model of supply-determined growth is that implicit in the traditional neoclassical analysis. The mechanism that adjusts demand to growing supply is the price mechanism, or Adam Smith's "invisible hand" of the market. This type of model assumes a world devoid of monopoly and uncertainty, in which the markets for capital goods and labour are free to adjust quickly so that "markets are always cleared" in the very short run.

A final example of a model of growth that illustrates the problem of adjustment between supply and demand is to be found in the work of the Dutch economist Jan Tinbergen and his followers. In contrast to neoclassical growth models where the market brings about an adjustment of demand to supply, the "target-instrument" models of Tinbergen assume that the government (as in The Netherlands and other European countries) undertakes to regulate demand and supply in an effort to achieve certain targets such as full employment or a predetermined rate of growth. For example, economists are expected to provide the fiscal authorities with a model that approximates the working of the economy and that indicates what will

happen if the government, say, does not change its tax and spending programs in the coming period. These forecasts are appraised in terms of what the authorities consider desirable as a matter of social and economic policy. If it appears that unemployment will be too high and the rate of growth too low, the authorities take countermeasures. The government may, for example, cut taxes on corporate profits in order to stimulate investment. If investment is excessive and there is danger of inflation, the government may take other measures to reduce aggregate demand such as cutting its expenditures. This type of planning procedure has been tried with varying degrees of success. Sweden and The Netherlands are prominent examples of attempts to offset fluctuations in private spending so as to realize full employment and growth. It should be noted that these models do not fit neatly into the demand-determined or supply-determined classification. In the example just given, both the rate of growth of demand and the rate of growth of supply are effectively determined by the fiscal authorities.

Economic stagnation.

The rise in unemployment rates and the slowdown in growth rates of GNP and per capita incomes throughout the capitalist world beginning in the early 1970s is clearly a case where demand and supply did not grow at similar rates. Many economists turned their attention to developing theories to explain this prolonged period of stagnation. A common theme in much of their work was the adverse effects of high unemployment and low utilization of the capital stock on investment and, therefore, on productivity growth.

The high unemployment rates for labour and capital are initially traced to policies restricting aggregate demand that were pursued by monetary and fiscal authorities from the first half of the 1970s. This policy response was widely interpreted by economists as an effort by the authorities to reduce inflation rates that had begun to accelerate in the latter 1960s. The continued use of restrictive policies is then related to fear on the part of the authorities that any attempt to restimulate their economies would merely bring back inflation.

Tighter labour markets resulting from any such stimulative policies are seen to increase the bargaining power of labour, thereby leading to larger wage demands and settlements that in turn feed into prices, causing price inflation to accelerate. This leads to yet higher wage demands in order to protect real wages and thus an explosive wage-price spiral. In addition, more stimulative aggregate demand policies are perceived to result in balance of payments difficulties at existing exchange rates. But any attempt to avoid larger payments deficits by reducing the exchange rate leads to the "importation" of inflation through higher prices of imported goods. The result of such considerations is reluctance of the authorities to attempt to create full employment through stimulative policies.

What emerges from these theories is a chain of causation that describes the way in which, in the period since World War II, inflation and growth have become causally connected through the responses of governments to actual and anticipated inflationary pressures. Inflation and the fear of inflation lead to slow growth and high unemployment because the inability of governments to bring inflation under control at full employment by other means - e.g., an income policy - constrains governments to implement restrictive policies to combat or forestall inflationary pressures. Such responses lead, as they did in the early 1970s, not only to high rates of unemployment of capital and labour but also to low rates of investment and productivity growth. Stagnation is the result, and such a scenario is a likely prospect for capitalism in the future.

Foreign trade.

Little has been said about foreign trade. Yet growth in most economies is very much dependent upon imports and the ability to export in order to pay for imports. The fact that some economies recovered relatively quickly from World War II and grew much more rapidly in the postwar period than others has stimulated a great deal of comparative analysis in growth theory. The exceptionally high growth rates in Japan and Germany compared to the general sluggishness of the British economy are related to foreign trade. Economists have pointed to the periodic balance of payments crises experienced by Britain and the lack of such crises in

Germany. During a boom, as incomes rise the demand for imports will rise also as a natural feature of prosperity. But if exports do not also rise at the same time, the authorities may be forced to take fiscal or monetary countermeasures and slow down the economy in an effort to bring imports and exports back into balance. Or exports may fail to grow sufficiently because labour costs are rising very rapidly and pushing up prices of exports faster than in competing countries.

A policy of encouraging growth has the effect of keeping the demand for imports high and making labour markets tight, thereby tending to push up money wage rates. At the same time, such a policy also tends to encourage innovations and investment projects that are very productive, particularly if the demand pressures are sustained. A "stop" policy naturally has just the opposite effects, both good and bad from the point of view of a country's balance of payments. The question is which policy will in the long run result in less rapidly rising costs and prices. Many writers have argued that if demand pressures are maintained the response or adjustment of productivity and therefore of supply to these pressures will be such that the country will soon find itself in a more competitive position. Running an economy "flat out," however, is likely to cause a short-run balance of payments crisis and lead to devaluation of currency.

Mathematical growth theories.

In addition to the theories discussed above, a large body of literature has developed involving abstract mathematical models. Because this field of analysis is so technical, only a general picture of the kinds of problems and questions discussed can be given. First, a set of equations is drawn up describing what the model builder feels are the important relations between economic variables such as output, capital, investment, and consumption. These equations must relate economic variables to one another at different points in time: for example, output last year determines consumption this year, which in turn helps to determine output this year and therefore consumption and output next year. It is possible to work out the movements of the variables over as long a period as desired. At the centre of much of this analysis is the concept of a steady-state rate of growth: one in which

all the economic variables contained in the set of equations grow at the same constant rate equal to the growth of the labour force.

A related class of studies attempts to take account of the welfare of workers and consumers in the maximization of growth. These "optimal growth" models seek to maximize consumer satisfaction over time. In a model such as this the solution will not be the highest possible growth rate but one that will maximize the welfare of consumers. The importance of such models for planners would seem to depend on the realism of their assumptions as to consumer desires and technology.

Model building and theorizing about growth has proceeded on various levels of abstraction. Some of the work is of little practical value, in the sense that its explanatory value is negligible. Such studies, however, may stimulate other work that is helpful in an understanding of the growth process. Some models, while realistic, are not applicable to all economies. Thus, a model that neglects international trade is of little use to a European economist trying to understand the more basic causes of differences in growth rates between countries.

Economic development

Economic development first became a major concern after World War II. As the era of European colonialism ended, many former colonies and other countries with low living standards came to be termed underdeveloped countries, to contrast their economies with those of the developed countries, which were understood to be Canada, the United States, those of western Europe, most eastern European countries, the then Soviet Union, Japan, South Africa, Australia, and New Zealand. As living standards in most poor countries began to rise in subsequent decades, they were renamed the developing countries.

There is no universally accepted definition of what a developing country is; neither is there one of what constitutes the process of economic development. Developing countries are usually categorized by a per capita income criterion, and economic development is usually thought to occur as per capita incomes rise. A country's per capita income (which is almost synonymous with per capita output) is the best available measure of the value of the goods and services available, per person, to

the society per year. Although there are a number of problems of measurement of both the level of per capita income and its rate of growth, these two indicators are the best available to provide estimates of the level of economic well-being within a country and of its economic growth.

It is well to consider some of the statistical and conceptual difficulties of using the conventional criterion of underdevelopment before analyzing the causes of underdevelopment. The statistical difficulties are well known. To begin with, there are the awkward borderline cases. Even if analysis is confined to the underdeveloped and developing countries in Asia, Africa, and Latin America, there are rich oil countries that have per capita incomes well above the rest but that are otherwise underdeveloped in their general economic characteristics. Second, there are a number of technical difficulties that make the per capita incomes of many underdeveloped countries (expressed in terms of an international currency, such as the U.S. dollar) a very crude measure of their per capita real income. These difficulties include the defectiveness of the basic national income and population statistics, the inappropriateness of the official exchange rates at which the national incomes in terms of the respective domestic currencies are converted into the common denominator of the U.S. dollar, and the problems of estimating the value of the noncash components of real incomes in the underdeveloped countries. Finally, there are conceptual problems in interpreting the meaning of the international differences in the per capita income levels.

Although the difficulties with income measures are well established, measures of per capita income correlate reasonably well with other measures of economic well-being, such as life expectancy, infant mortality rates, and literacy rates. Other indicators, such as nutritional status and the per capita availability of hospital beds, physicians, and teachers, are also closely related to per capita income levels. While a difference of, say, 10 percent in per capita incomes between two countries would not be regarded as necessarily indicative of a difference in living standards between them, actual observed differences are of a much larger magnitude. India's per capita income, for example, was estimated at $270 in 1985. In contrast, Brazil's

was estimated to be $1,640, and Italy's was $6,520. While economists have cited a number of reasons why the implication that Italy's living standard was 24 times greater than India's might be biased upward, no one would doubt that the Italian living standard was significantly higher than that of Brazil, which in turn was higher than India's by a wide margin.

The interpretation of a low per capita income level as an index of poverty in a material sense may be accepted with two qualifications. First, the level of material living depends not on per capita income as such but on per capita consumption. The two may differ considerably when a large proportion of the national income is diverted from consumption to other purposes; for example, through a policy of forced saving. Second, the poverty of a country is more faithfully reflected by the representative standard of living of the great mass of its people. This may be well below the simple arithmetic average of per capita income or consumption when national income is very unequally distributed and there is a wide gap in the standard of living between the rich and the poor.

The usual definition of a developing country is that adopted by the World Bank: "low-income developing countries" in 1985 were defined as those with per capita incomes below $400; "middle-income developing countries" were defined as those with per capita incomes between $400 and $4,000. To be sure, countries with the same per capita income may not otherwise resemble one another: some countries may derive much of their incomes from capital-intensive enterprises, such as the extraction of oil, whereas other countries with similar per capita incomes may have more numerous and more productive uses of their labour force to compensate for the absence of wealth in resources. Kuwait, for example, was estimated to have a per capita income of $14,480 in 1985, but 50 percent of that income originated from oil. In most regards, Kuwait's economic and social indicators fell well below what other countries with similar per capita incomes had achieved. Centrally planned economies are also generally regarded as a separate class, although China and North Korea are universally considered developing countries. A major difficulty is that prices serve less as indicators of relative scarcity in centrally

planned economies and hence are less reliable as indicators of the per capita availability of goods and services than in market-oriented economies.

Estimates of percentage increases in real per capita income are subject to a somewhat smaller margin of error than are estimates of income levels. While year-to-year changes in per capita income are heavily influenced by such factors as weather (which affects agricultural output, a large component of income in most developing countries), a country's terms of trade, and other factors, growth rates of per capita income over periods of a decade or more are strongly indicative of the rate at which average economic well-being has increased in a country.

ECONOMIC DEVELOPMENT AS AN OBJECTIVE OF POLICY

Motives for development.

The field of development economics is concerned with the causes of underdevelopment and with policies that may accelerate the rate of growth of per capita income. While these two concerns are related to each other, it is possible to devise policies that are likely to accelerate growth (through, for example, an analysis of the experiences of other developing countries) without fully understanding the causes of underdevelopment.

Studies of both the causes of underdevelopment and of policies and actions that may accelerate development are undertaken for a variety of reasons. There are those who are concerned with the developing countries on humanitarian grounds; that is, with the problem of helping the people of these countries to attain certain minimum material standards of living in terms of such factors as food, clothing, shelter, and nutrition. For them, low per capita income is the measure of the problem of poverty in a material sense. The aim of economic development is to improve the material standards of living by raising the absolute level of per capita incomes. Raising per capita incomes is also a stated objective of policy of the governments of all developing countries. For policymakers and economists attempting to achieve their governments' objectives, therefore, an understanding of economic development, especially in its policy dimensions, is important. Finally, there are those who are concerned with economic development either because they

believe it is what people in developing countries want or because they believe that political stability can be assured only with satisfactory rates of economic growth. These motives are not mutually exclusive. Since World War II many industrial countries have extended foreign aid to developing countries for a combination of humanitarian and political reasons.

Those who are concerned with political stability tend to see the low per capita incomes of the developing countries in relative terms; that is, in relation to the high per capita incomes of the developed countries. For them, even if a developing country is able to improve its material standards of living through a rise in the level of its per capita income, it may still be faced with the more intractable subjective problem of the discontent created by the widening gap in the relative levels between itself and the richer countries. (This effect arises simply from the operation of the arithmetic of growth on the large initial gap between the income levels of the developed and the underdeveloped countries. As an example, an underdeveloped country with a per capita income of $100 and a developed country with a per capita income of $1,000 may be considered. The initial gap in their incomes is $900. Let the incomes in both countries grow at 5 percent. After one year, the income of the underdeveloped country is $105, and the income of the developed country is $1,050. The gap has widened to $945. The income of the underdeveloped country would have to grow by 50 percent to maintain the same absolute gap of $900.) Although there was once in development economics a debate as to whether raising living standards or reducing the relative gap in living standards was the true desideratum of policy, experience during the 1960-80 period convinced most observers that developing countries could, with appropriate policies, achieve sufficiently high rates of growth both to raise their living standards fairly rapidly and to begin closing the gap.

The impact of discontent.

Although concern over the question of a subjective sense of discontent among the underdeveloped and developing countries has waxed and waned, it has never wholly disappeared. The underdeveloped countries' sense of dissatisfaction and

grievance arises not only from measurable differences in national incomes but also from the less easily measurable factors, such as their reaction against the colonial past and their complex drives to raise their national prestige and achieve equality in the broadest sense with the developed countries. Thus, it is not uncommon to find their governments using a considerable proportion of their resources in prestige projects, ranging from steel mills, hydroelectric dams, universities, and defense expenditure to international athletics. These symbols of modernization may contribute a nationally shared satisfaction and pride but may or may not contribute to an increase in the measurable national income. Second, it is possible to argue that in many cases the internal gap in incomes within individual underdeveloped countries may be a more potent source of the subjective level of discontent than the international gap in income. Faster economic growth may help to reduce the internal economic disparities in a less painful way, but it must be remembered that faster economic growth also tends to introduce greater disruption and the need for making bigger readjustments in previous ways of life and may thus increase the subjective sense of frustration and discontent. Finally, it is difficult to establish that the subjective problem of discontent will bear a simple and direct relationship to the size of the international gap in incomes. Some of the apparently most discontented countries are to be found in Latin America, where the per capita incomes are generally higher than in Asia and Africa. A skeptic can turn the whole approach to a reductio ad absurdum by pointing out that even the developed countries with their high and rising levels of per capita income have not been able to solve the subjective problem of discontent and frustration among various sections of their population.

Two conclusions may be drawn from the above points. First, the subjective problem of discontent in the underdeveloped countries is a genuine and important problem in international relations. But economic policy acting on measurable economic magnitudes can play only a small part in the solution of what essentially is a problem in international politics. Second, for the narrower purpose of economic policy there is no choice but to fall back on the interpretation of the low

per capita incomes of the underdeveloped countries as an index of their poverty in a material sense. This can be defended by explicitly adopting the humanitarian value judgment that the underdeveloped countries ought to give priority to improving the material standards of living of the mass of their people. But, even if this value judgment is not accepted, the conventional measure of economic development in terms of a rise in per capita income still retains its usefulness. The governments of the underdeveloped countries may wish to pursue other, nonmaterial goals, but they could make clearer decisions if they knew the economic cost of their decisions. The most significant measure of this economic cost can be expressed in terms of the foregone opportunity to raise the level of per capita income.

A SURVEY OF DEVELOPMENT THEORIES

The hypothesis of underdevelopment.

If the underdeveloped countries are merely low-income countries, why call them underdeveloped? The use of the term underdeveloped in fact rests on a general hypothesis on which the whole subject matter of development economics is based. According to this hypothesis, the existing differences in the per capita income levels between the developed and the underdeveloped countries cannot be accounted for purely in terms of differences in natural conditions beyond the control of man and society. That is to say, the underdeveloped countries are underdeveloped because, in some way or another, they have not yet succeeded in making full use of their potential for economic growth. This potential may arise from the underdevelopment of their natural resources, or their human resources, or from the "technological gap." More generally, it may arise from the underdevelopment of economic organization and institutions, including the network of the market system and the administrative machinery of the government. The general presumption is that the development of this organizational framework would enable an underdeveloped country to make a fuller use not only of its domestic resources but also of its external economic opportunities, in the form of

international trade, foreign investment, and technological and organizational innovations.

Development thought after World War II.

After World War II a number of developing countries attained independence from their former colonial rulers. One of the common claims made by leaders of independence movements was that colonialism had been responsible for perpetuating low living standards in the colonies. Thus economic development after independence became an objective of policy not only because of the humanitarian desire to raise living standards but also because political promises had been made, and failure to make progress toward development would, it was feared, be interpreted as a failure of the independence movement. Developing countries in Latin America and elsewhere that had not been, or recently been, colonies took up the analogous belief that economic domination by the industrial countries had thwarted their development, and they, too, joined the quest for rapid growth.

At that early period, theorizing about development, and about policies to attain development, accepted the assumption that the policies of the industrial countries were to blame for the poverty of the developing countries. Memories of the Great Depression, when developing countries' terms of trade had deteriorated markedly, producing sharp reductions in per capita incomes, haunted many policymakers. Finally, even in the developed countries, the Keynesian legacy attached great importance to investment.

In this milieu, it was thought that a "shortage of capital" was the cause of underdevelopment. It followed that policy should aim at an accelerated rate of investment. Since most countries with low per capita incomes were also heavily agricultural (and imported most of the manufactured goods consumed domestically), it was thought that accelerated investment in industrialization and the development of manufacturing industries to supplant imports through "import substitution" was the path to development. Moreover, there was a fundamental distrust of markets, and a major role was therefore assigned to government in

allocating investments. Distrust of markets extended especially to the international economy.

Experience with development changed perceptions of the process and of the policies affecting it in important ways. Nonetheless, there are significant elements of truth in some of the earlier ideas, and it is important to understand the thinking underlying them.

Growth economics and development economics.

Development economics may be contrasted with another branch of study, called growth economics, which is concerned with the study of the long-run, or steady-state, equilibrium growth paths of the economically developed countries, which have long overcome the problem of initiating development.

Growth theory assumes the existence of a fully developed modern capitalist economy with a sufficient supply of entrepreneurs responding to a well-articulated system of economic incentives to drive the growth mechanism. Typically, it concentrates on macroeconomic relations, particularly the ratio of savings to total output and the aggregate capital-output ratio (that is, the number of units of additional capital required to produce an additional unit of output). Mathematically, this can be expressed (the Harrod-Domar growth equation) as follows: the growth in total output (g) will be equal to the savings ratio (s) divided by the capital-output ratio (k); i.e., $g = s/k$. Thus, suppose that 12 percent of total output is saved annually and that three units of capital are required to produce an additional unit of output: then the rate of growth in output is 12/3% = 4% per annum. This result is obtained from the basic assumption that whatever is saved will be automatically invested and converted into an increase in output on the basis of a given capital-output ratio. Since a given proportion of this increase in output will be saved and invested on the same basis, a continuous process of growth is maintained.

Growth theory, particularly the Harrod-Domar growth equation, has been frequently applied or misapplied to the economic planning of a developing country. The planner starts from a desired target rate of growth of perhaps 4

percent. Assuming a fixed overall capital-output ratio of, say, 3, it is then asserted that the developing country will be able to achieve this target rate of growth if it can increase its savings to 3 4 percent = 12 percent of its total output. The weakness of this type of exercise arises from the assumption of a fixed overall capital-output ratio, which assumes away all the vital problems affecting the developing country's capacity to absorb capital and invest its saving in a productive manner. These problems include the central problem of the efficient allocation of available savings among alternative investment opportunities and the associated organizational and institutional problems of encouraging the growth of a sufficient supply of entrepreneurs; the provision of appropriate economic incentives through a market system that correctly reflects the relative scarcities of products and factors of production; and the building up of an organizational framework that can effectively implement investment decisions in both the private and the public sectors. Such problems, which generally affect the developing country's absorptive capacity for capital and a number of other inputs, constitute the core of development economics. Development economics is needed precisely because the assumptions of growth economics, based as they are on the existence of a fully developed and well-functioning modern capitalist economy, do not apply.

The developing and underdeveloped countries are a very mixed collection of countries. They differ widely in area, population density, and natural resources. They are also at different stages in the development of market and financial institutions and of an effective administrative framework. These differences are sufficient to warn against wide-sweeping generalizations about the causes of underdevelopment and all-embracing theoretical models of economic development. But when development economics first came into prominence in the 1950s, there were powerful intellectual and political forces propelling the subject toward such general theoretical models of development and underdevelopment. First, many writers who popularized the subject were frankly motivated by a desire to persuade the developed countries to give more economic aid to the underdeveloped countries, on grounds ranging from humanitarian considerations to

considerations of cold-war strategy. Second, there was the reaction of the newly independent underdeveloped countries against their past "colonial economic pattern," which they identified with free trade and primary production for the export market. These countries were eager to accept general theories of economic development that provided a rationalization for their deep-seated desire for rapid industrialization. Third, there was a parallel reaction, at the academic level, against older economic theory, with its emphasis on the efficient allocation of scarce resources and a striving after new and "dynamic" approaches to economic development.

All of these forces combined to produce a crop of theoretical approaches that soon developed into a fairly fixed orthodoxy with its characteristic emphasis on "crash" programs of investment in both material and human capital, on domestic industrialization, and on government economic planning as the standard ingredients of development policy. These new theories have continued to have a considerable influence on the conventional wisdom in development economics, although in retrospect most of them have turned out to be partial theories. A broad survey of these theories, under three main heads, is given below. It is particularly relevant to the debate over whether the underdeveloped countries should seek economic development through domestic industrialization or through international trade. The limitations of these new theories - and how they led to a gradual revival of a more pragmatic approach todevelopment problems, which falls back increasingly on the older economic theory of efficient allocation of resources - are subsequently traced.

The missing-component approach.

First, there are the theories that regard the shortage of some strategic input (such as the supply of savings, foreign exchange, or technical skills) as the main cause of underdevelopment. Once this missing component was supplied - say, by external economic aid - it was believed that economic development would follow in a predictable manner based on fixed quantitative relationships between input and output. The overall capital-output ratio, mentioned above, is the most well-known

of these fixed technical coefficients. But similar fixed coefficients have been assumed between the foreign-exchange requirements and total output and between the input of skilled manpower and output.

Shortage of savings.

Given the broad relationship between capital accumulation and economic growth established in growth theory, it was plausible for growth theorists and development economists to argue that the developing countries were held back mainly by a shortage in the supply of capital. These countries were then saving only 5-7 percent of their total product, and it was manifest (and it remains true) that satisfactory growth cannot be supported by so low a level of investment. It was therefore thought that raising the savings ratio to 10-12 percent was the central problem for developing countries. Early development policy therefore focused on raising resources for investment. Steps toward this end were highly successful in most developing countries, and savings ratios rose to the 15-25 percent range. However, growth rates failed even to approximate the savings rates, and theorists were forced to search for other explanations of differences in growth rates.

It has become increasingly clear that there can be much wastage of capital resources in the developing countries for various reasons, such as wrong choice of investment projects, inefficient implementation and management of these projects, and inappropriate pricing and costing of output. These faults are particularly noticeable in public-sector investment projects and are one of the reasons why the Pearson Commission Report of the International Bank for Reconstruction and Development (1969) found that "the correlation between the amounts of aid received in the past decades and the growth performance is very weak." But even in the private sector there may be a considerable distortion in the direction of investment induced by policies designed to encourage development. Thus, in most underdeveloped countries, a considerable part of private expansion investment, both foreign and domestic, has been diverted into the expansion of the manufacturing sector, catering to the domestic market through various inducements, including tariff protection, tax holidays, cheap loans, and generous

foreign-exchange allocations granting the opportunity to import capital goods cheaply at overvalued exchange rates. As a consequence, there developed a very considerable amount of excess capacity in the manufacturing sector of the underdeveloped countries pursuing such policies.

Foreign-exchange shortage.

In the 1950s most developing countries were primary commodity exporters, relying on crops and minerals for the bulk of their foreign-exchange earnings through exports, and importing a large number of manufactured goods. The experience of colonialism, and the distrust of the international economy that it engendered, led policymakers in most developing countries to adopt a policy of import substitution. This policy was intended to promote industrialization by protecting domestic producers from the competition of imports. Protection, in the form of high tariffs or the restriction of imports through quotas, was applied indiscriminately, often to inherently high-cost industries that had no hope of ever becoming internationally competitive. Also, after the early stages of import substitution, protected new industries tended to be very intensive in the use of capital and especially of imported capital goods.

The import-substitution approach defined "industrialization" rather narrowly as the expansion of the modern manufacturing sector based on capital-intensive technology. Capital was therefore identified with durable capital equipment in the form of complex machinery and other inputs that the underdeveloped countries were not able to produce domestically. Thus, foreign-exchange requirements were calculated on the basis of the fixed technical input-output coefficients of the manufacturing sector.

With high levels of protection for domestic industry, and with exchange rates that were often maintained at unrealistic levels (usually in an effort to make imported capital goods "cheap"), the experience of most developing countries was that export earnings grew relatively slowly. The simultaneously sharp increase in demand for imported capital goods (and for raw materials and replacement parts as well) resulted in unexpectedly large increases in imports. Most developing

countries found themselves with critical foreign-exchange shortages and were forced to reduce imports in order to cut their current-account deficits to manageable proportions.

The cutbacks in imports usually resulted in reduced growth rates, if not recessions. This result led to the view that economic stagnation was caused primarily by a shortage of foreign exchange with which to buy essential industrial inputs. But over the longer term the growth rates of countries that continued to protect their domestic industries heavily not only stagnated but declined sharply. Contrasting the experience of countries that persisted in policies of import substitution with those that followed alternative policies (see below) subsequently demonstrated that foreign-exchange shortage was a barrier to growth only within the context of the protectionist policies adopted and was not inherently a barrier to the development process itself.

Education and human capital in development.

As it became apparent that the physical accumulation of capital was not by itself the key to development, many analysts turned to a lack of education and skills among the population as being a crucial factor in underdevelopment. If education and skill are defined as everything that is required to raise the productivity of the people in the developing countries by improving their skills, enterprise, initiative, adaptability, and attitudes, this proposition is true but is an empty tautology. However, the need for skills and training was first formulated in terms of specific skills and educational qualifications that could be supplied by crash programs in formal education. The usual method of manpower planning thus started from a target rate of expansion in output and tried to estimate the numbers of various types of skilled personnel that would be required to sustain this target rate of economic growth on the basis of an assumed fixed relationship between inputs of skill and national output.

This approach was plausible enough in many developing countries immediately after their political independence, when there were obvious gaps in various branches of the administrative and technical services. But most countries passed

through this phase rather quickly. In the meantime, as the result of programs in education expansion, their schools and colleges began producing large numbers of fresh graduates at much faster rates than their general rate of economic growth could supply suitable new jobs for. This created a growing problem of educated unemployment. An important factor behind the rapid educational expansion was the expectation that after graduation students would be able to obtain well-paying white-collar jobs at salary levels many times the prevailing per capita income of their countries. Thus, the underdeveloped countries' inability to create jobs to absorb their growing armies of graduates created an explosive element in what came to be called the revolution of expectations.

It is possible to see a close parallelism between the narrow concept of industrialization as the expansion of the manufacturing sector and the narrow concept of education as the academic and technical qualifications that can be supplied by the expansion of the formal educational system. If a broader concept of education, relevant for economic development, is needed, it is necessary to seek it in the pervasive educational influence of the economic environment as a whole on the learning process of the people of the underdeveloped countries. This is a complex process that depends on, among other less easily analyzable things, the system of economic incentives and signals that can mold the economic behaviour of the people of the underdeveloped countries and affect their ability to make rational economic decisions and their willingness to introduce or adapt to economic changes. Unfortunately, the economic environment in many underdeveloped countries is dominated by a network of government controls that tend not to be conducive to such ends.

Surplus resources and disguised unemployment.

Two theories emphasized the existence of surplus resources in developing countries as the central challenge for economic policy. The first concentrated on the countries with relatively abundant natural resources and low population densities and argued that a considerable amount of both surplus land and surplus labour might still exist in these countries because of inadequate marketing facilities

and lack of transport and communications. Economic development was pictured as a process whereby these underutilized resources of the subsistence sector would be drawn into cash production for the export market. International trade was regarded as the chief market outlet, or vent, for the surplus resources. The second theory was concerned with the thickly populated countries and the possibility of using their surplus labour as the chief means of promoting economic development. According to this theory, because of heavy population pressure on land, the marginal product of labour (that is, the extra output derived from the employment of an extra unit of labour) was reduced to zero or to a very low level. But the people in the subsistence sector were able to enjoy a certain customary minimum level of real income because the extended-family system of the rural society shared the total output of the family farm among its members. A considerable proportion of labour in the traditional agricultural sector was thus thought to contribute little or nothing to total output and to really be in a state of disguised unemployment. By this theory, the labour might be drawn into other uses without any cost to society.

It is necessary to clear up a number of preliminary points about the concept of disguised unemployment before considering its applications. First, it is highly questionable whether the marginal product of labour is actually zero even in densely populated countries such as India or Pakistan. Even in these countries, with existing agricultural methods, all available labour is needed in the peak seasons, such as harvest. The most important part of disguised unemployment is thus what may be better described as seasonal unemployment during the off-seasons. The magnitude of this seasonal unemployment, however, depends not so much on the population density on land as on the number of crops cultivated on the same piece of land through the year. There is thus little seasonal unemployment in countries such as Taiwan or South Korea, which have much higher population densities than India, because improved irrigation facilities enable them to grow a succession of crops on the same land throughout the year. But there may be considerable seasonal unemployment even in sparsely populated countries growing only one crop a year.

The main weakness in the proposal to use disguised unemployment for the construction of major social-overhead-capital projects arises from an inadequate consideration of the problem of providing the necessary subsistence fund to maintain the workers during what may be a considerably long waiting period before these projects yield consumable output. This may be managed somehow for small-scale local-community projects when the workers are maintained in situ by their relatives. But when it is proposed to move a large number of surplus workers away from their home villages for major construction projects taking a considerable time to complete, the problem of raising a sufficient subsistence fund to maintain the labour becomes formidable. The only practicable way of raising such a subsistence fund is to encourage voluntary saving and the expansion of a marketable surplus of food that can be purchased with the savings to maintain the workers. The mere existence of disguised unemployment does not in any way ease this problem.

Role of governments and markets.

In earlier thinking about development, it was assumed that the market mechanisms of developed economies were so unreliable in developing economies that governments had to assume central responsibility for economic activity. This was to be done through economic planning for the entire economy (see below Planning in developing countries), which in turn would be implemented by active government participation in the economy and pervasive controls over all private-sector economic activity. Government participation took many forms: Public-sector enterprises were established to manufacture many commodities, including steel, machine tools, fertilizers, heavy chemicals, and even textiles and clothing; government marketing boards assumed monopoly power over the purchase and sale of many agricultural commodities; and government agencies became the sole importers of a variety of goods, and they often became exporters as well. Controls over private-sector activity were even more extensive: Price controls were established for many commodities; import licensing procedures eliminated the importing of commodities not given priority in official plans; investment licenses

were required before factories could be expanded; capacity licenses regulated maximum permissible outputs; and comprehensive regulations governed the conditions of employment of workers.

The consequence, frequently, was that indigenous entrepreneurs often found it more financially rewarding to devote their energies and ingenuity to the task of procuring the necessary government import licenses and other permits and exploiting the loopholes in government regulations than to the problem of raising the efficiency and productivity of resources. For public-sector enterprises, political pressures often resulted in the employment of many more persons than could be productively used and in other practices conducive to extremely high-cost and inefficient operations. The consequent fiscal burden diverted resources that might otherwise have been used for investment, while the inefficient use of resources dampened growth rates.

Related to the belief in market failure and in the necessity for government intervention was the view that the efficiency of the price mechanism in developing countries was very small. This was reflected in the view of foreign-exchange shortage, already discussed, in which it was thought that there are fixed relationships between imported capital and domestic expansion. It was also reflected in the view that farmers are relatively insensitive to relative prices and in the belief that there are few entrepreneurs in developing countries.

ФАКУЛЬТЕТ КУЛЬТУРЫ И ИСКУССТВА

LANGUAGE CLASSIFICATION

There are two kinds of classification of languages practiced in linguistics: genetic (or genealogical) and typological. The purpose of genetic classification is to group languages into families according to their degree of diachronic relatedness. For example, within the Indo-European family, such subfamilies as Germanic or Celtic are recognized; these subfamilies comprise German, English, Dutch, Swedish, Norwegian, Danish, and others, on the one hand, and Irish, Welsh, Breton, and others, on the other. So far, most of the languages of the world have been grouped only tentatively into families, and many of the classificatory schemes that have been proposed will no doubt be radically revised as further progress is made.

A typological classification groups languages into types according to their structural characteristics. The most famous typological classification is probably that of isolating, agglutinating, and inflecting (or fusional) languages, which was frequently invoked in the 19th century in support of an evolutionary theory of language development. Roughly speaking, an isolating language is one in which all the words are morphologically unanalyzable (i.e., in which each word is composed of a single morph); Chinese and, even more strikingly, Vietnamese are highly isolating. An agglutinating language (e.g., Turkish) is one in which the word forms can be segmented into morphs, each of which represents a single grammatical category. An inflecting language is one in which there is no one-to-one correspondence between particular word segments and particular grammatical categories. The older Indo-European languages tend to be inflecting in this sense. For example, the Latin suffix -is represents the combination of categories "singular" and "genitive" in the word form hominis "of the man," but one part of the suffix cannot be assigned to "singular" and another to "genitive," and -is is only one of many suffixes that in different classes (or declensions) of words represent the combination of "singular" and "genitive."

There is, in principle, no limit to the variety of ways in which languages can be grouped typologically. One can distinguish languages with a relatively rich

phonemic inventory from languages with a relatively poor phonemic inventory, languages with a high ratio of consonants to vowels from languages with a low ratio of consonants to vowels, languages with a fixed word order from languages with a free word order, prefixing languages from suffixing languages, and so on. The problem lies in deciding what significance should be attached to particular typological characteristics. Although there is, not surprisingly, a tendency for genetically related languages to be typologically similar in many ways, typological similarity of itself is no proof of genetic relationship. Nor does it appear true that languages of a particular type will be associated with cultures of a particular type or at a certain stage of development. What has emerged from recent work in typology is that certain logically unconnected features tend to occur together, so that the presence of feature A in a given language will tend to imply the presence of feature B. The discovery of unexpected implications of this kind calls for an explanation and gives a stimulus to research in many branches of linguistics.

Meaning and use.

The difficulties just mentioned lead to another view concerning the notion of meaning, a theory that may be called the use theory. This view admits that not all words refer to something, and not all utterances are true or false. What is common to all words and all sentences, without exception, is that people use them in speech. Consequently, their meaning may be nothing more than the restrictions, rules, and regularities that govern their employment.

The use theory has several sources. First, in trying to understand the nature of moral and aesthetic discourse certain authors suggested that such words as "good" and "beautiful" have an emotive meaning instead of (or in addition to) the descriptive meaning other words have; in using them one expresses approval or commendation. If one says, for instance, that helping the poor is good, one does not describe that action, but says, in effect, something like "I approve of helping the poor, do so as well." Such is the role of these words, according to these thinkers, and to understand this role is to know their meaning.

The second, and more important, stimulus for the use theory was provided by the work of Ludwig Wittgenstein. This philosopher not only pointed out the wide variety of linguistic moves mentioned above but in order to show that none of these moves enjoys a privileged status proposed the idea of certain language games in which one or another of these moves plays a dominant or even an exclusive role. One can imagine, for instance, a tribe whose language consists of requests only. Members of the tribe make requests and the other members comply or refuse. There is no truth in this language, yet the words used to make requests would have meaning. Human language as it exists in reality is more complex; it is a combination of a great many language games. Yet the principle of meaning, according to this theory, is the same: the meaning of a word is the function of its employment in these games. To Wittgenstein the question "What is a word really?" is analogous to "What is a piece in chess?" Finally, John L. Austin offered a systematic classification of the variety of speech acts. According to him, to say something is to do something, and what one does in saying something is typically indicated by a particular performative verb prefixing the "normal form" of the utterance. These verbs, such as "state," "declare," "judge," "order," "request," "promise," "warn," "apologize," "call," and so on, mark the illocutionary force of the utterance in question. If one says, for instance, "I shall be there," then, depending on the circumstances, this utterance may amount to a prediction, a promise, or a warning. Similarly, the words of the commanding officer, "You will retreat" may have the force of a simple forecast, or of an order. If the circumstances are not clear, the speaker always can be more explicit and use the normal form; e.g., "I promise that I shall be there" or "I order you to retreat."

To rephrase the conclusion already stated: the dimension of truth and falsity is not invoked by all the utterances of the language; therefore, it cannot provide an exclusive source of meaning. There are other dimensions, such as feasibility (in case of orders and promises), utility (in case of regulations and prescriptions), and moral worth (in case of advices and laws). These dimensions may be as much

involved in the understanding of what one said and, consequently, in the meaning of the words the speaker used, as the dimension of truth.

As previously mentioned, philosophers professing the aletheic theory claimed that the meaning of a word should be explained in terms of its contribution to the truth or falsity of the sentences in which it can occur. The latest form of the use theory is an appropriate extension of the same idea. According to some exponents, the meaning of a word is nothing but its illocutionary act potential - i.e., its contribution to the nature of the speech acts that can be performed by using that word. One difficulty with this view is that the definition is too broad, to the extent of being unilluminating or useless. Given this definition, nobody would know what any word means without knowing the entire language completely because the possibilities of employing a given word are not only without limit but extend to every conceivable context and circumstance. As Wittgenstein stated so forcefully, The sign (the sentence) gets its significance from the system of signs, from the language to which it belongs. Roughly: understanding a sentence means understanding a language.

If this be the case, how can one account for the obviously gradual and prolonged process of learning a language? Indeed, the definition of the meaning of a word as illocutionary act potential seems to overstate the case. The obvious truth that the meaning of performative verbs, and other words closely tied to one illocutionary aspect or other, cannot be divorced from the nature of that type of speech act, does not entail that the meaning of an ordinary word, like "cat" or "running" is affected by any illocutionary force. Such words can occur in utterances bearing all kinds of illocutionary forces, so the contribution of these forces, as it were, cancel out. Nevertheless, what remains is the fact that all words are used to say something, in one way or another. The use theory would put a strong emphasis on the word "used" in the previous sentence. The next, and final, approach to meaning would stress the word "say."

Morphology.

The grammatical description of many, if not all, languages is conveniently divided into two complementary sections: morphology and syntax. The relationship between them, as generally stated, is as follows: morphology accounts for the internal structure of words, and syntax describes how words are combined to form phrases, clauses, and sentences.

There are many words in English that are fairly obviously analyzable into smaller grammatical units. For example, the word "unacceptability" can be divided into un-, accept, abil-, and -ity (abil- being a variant of -able). Of these, at least three are minimal grammatical units, in the sense that they cannot be analyzed into yet smaller grammatical units - un-, abil-, and ity. The status of accept, from this point of view, is somewhat uncertain. Given the existence of such forms as accede and accuse, on the one hand, and of except, excede, and excuse, on the other, one might be inclined to analyze accept into ac- (which might subsequently be recognized as a variant of ad-) and -cept. The question is left open. Minimal grammatical units like un-, abil-, and -ity are what Bloomfield called morphemes; he defined them in terms of the "partial phonetic-semantic resemblance" holding within sets of words. For example, "unacceptable," "untrue," and "ungracious" are phonetically (or, phonologically) similar as far as the first syllable is concerned and are similar in meaning in that each of them is negative by contrast with a corresponding positive adjective ("acceptable," "true," "gracious"). This "partial phonetic-semantic resemblance" is accounted for by noting that the words in question contain the same morpheme (namely, un-) and that this morpheme has a certain phonological form and a certain meaning.

Bloomfield's definition of the morpheme in terms of "partial phonetic-semantic resemblance" was considerably modified and, eventually, abandoned entirely by some of his followers. Whereas Bloomfield took the morpheme to be an actual segment of a word, others defined it as being a purely abstract unit, and the term morph was introduced to refer to the actual word segments. The distinction between morpheme and morph (which is, in certain respects, parallel to the distinction between phoneme and phone) may be explained by means of an

example. If a morpheme in English is posited with the function of accounting for the grammatical difference between singular and plural nouns, it may be symbolized by enclosing the term plural within brace brackets. Now the morpheme [plural] is represented in a number of different ways. Most plural nouns in English differ from the corresponding singular forms in that they have an additional final segment. In the written forms of these words, it is either -s or -es (e.g., "cat" : "cats"; "dog" : "dogs"; "fish" : "fishes"). The word segments written -s or -es are morphs. So also is the word segment written -en in "oxen." All these morphs represent the same morpheme. But there are other plural nouns in English that differ from the corresponding singular forms in other ways (e.g., "mouse" : "mice"; "criterion" : "criteria"; and so on) or not at all (e.g., "this sheep" : "these sheep"). Within the post-Bloomfieldian framework no very satisfactory account of the formation of these nouns could be given. But it was clear that they contained (in some sense) the same morpheme as the more regular plurals.

Morphs that are in complementary distribution and represent the same morpheme are said to be allomorphs of that morpheme. For example, the regular plurals of English nouns are formed by adding one of three morphs on to the form of the singular: /s/, /z/, or /iz/ (in the corresponding written forms both /s/ and /z/ are written -s and /iz/ is written -es). Their distribution is determined by the following principle: if the morph to which they are to be added ends in a "sibilant" sound (e.g., s, z, sh, ch), then the syllabic allomorph /iz/ is selected (e.g., fish-es /fis-iz/, match-es /mac-iz/); otherwise the nonsyllabic allomorphs are selected, the voiceless allomorph /s/ with morphs ending in a voiceless consonant (e.g., cat-s /kat-s/) and the voiced allomorph /z/ with morphs ending in a vowel or voiced consonant (e.g., flea-s /fli-z/, dog-s /dog-z/). These three allomorphs, it will be evident, are in complementary distribution, and the alternation between them is determined by the phonological structure of the preceding morph.

Thus the choice is phonologically conditioned.

Very similar is the alternation between the three principal allomorphs of the past participle ending, /id/, /t/, and /d/, all of which correspond to the -ed of the written

forms. If the preceding morph ends with /t/ or /d/, then the syllabic allomorph /id/ is selected (e.g., wait-ed /weit-id/). Otherwise, if the preceding morph ends with a voiceless consonant, one of the nonsyllabic allomorphs is selected - the voiceless allomorph /t/ when the preceding morph ends with a voiceless consonant (e.g., pack-ed /pak-t/) and the voiced allomorph /d/ when the preceding morph ends with a vowel or voiced consonant (e.g., row-ed /rou-d/; tame-d /teim-d/). This is another instance of phonological conditioning. Phonological conditioning may be contrasted with the principle that determines the selection of yet another allomorph of the past participle morpheme. The final /n/ of show-n or see-n (which marks them as past participles) is not determined by the phonological structure of the morphs show and see. For each English word that is similar to "show" and "see" in this respect, it must be stated as a synchronically inexplicable fact that it selects the /n/ allomorph. This is called grammatical conditioning. There are various kinds of grammatical conditioning.

Alternation of the kind illustrated above for the allomorphs of the plural morpheme and the /id/, /d/, and /t/ allomorphs of the past participle is frequently referred to as morphophonemic. Some linguists have suggested that it should be accounted for not by setting up three allomorphs each with a distinct phonemic form but by setting up a single morph in an intermediate morphophonemic representation. Thus, the regular plural morph might be said to be composed of the morphophoneme /Z/ and the most common past-participle morph of the morphophoneme /D/. General rules of morphophonemic interpretation would then convert /Z/ and /D/ to their appropriate phonetic form according to context. This treatment of the question foreshadows, on the one hand, the stratificational treatment and, on the other, the generative approach, though they differ considerably in other respects.

An important concept in grammar and, more particularly, in morphology is that of free and bound forms. A bound form is one that cannot occur alone as a complete utterance (in some normal context of use). For example, -ing is bound in this sense, whereas wait is not, nor is waiting. Any form that is not bound is free. Bloomfield

based his definition of the word on this distinction between bound and free forms. Any free form consisting entirely of two or more smaller free forms was said to be a phrase (e.g., "poor John" or "ran away"), and phrases were to be handled within syntax. Any free form that was not a phrase was defined to be a word and to fall within the scope of morphology. One of the consequences of Bloomfield's definition of the word was that morphology became the study of constructions involving bound forms. The so-called isolating languages, which make no use of bound forms (e.g., Vietnamese), would have no morphology.

The principal division within morphology is between inflection and derivation (or word formation). Roughly speaking, inflectional constructions can be defined as yielding sets of forms that are all grammatically distinct forms of single vocabulary items, whereas derivational constructions yield distinct vocabulary items. For example, "sings," "singing," "sang," and "sung" are all inflectional forms of the vocabulary item traditionally referred to as "the verb to sing"; but "singer," which is formed from "sing" by the addition of the morph -er (just as "singing" is formed by the addition of -ing), is one of the forms of a different vocabulary item. When this rough distinction between derivation and inflection is made more precise, problems occur. The principal consideration, undoubtedly, is that inflection is more closely integrated with and determined by syntax. But the various formal criteria that have been proposed to give effect to this general principle are not uncommonly in conflict in particular instances, and it probably must be admitted that the distinction between derivation and inflection, though clear enough in most cases, is in the last resort somewhat arbitrary.

Bloomfield and most linguists have discussed morphological constructions in terms of processes. Of these, the most widespread throughout the languages of the world is affixation; i.e., the attachment of an affix to a base. For example, the word "singing" can be described as resulting from the affixation of -ing to the base sing. (If the affix is put in front of the base, it is a prefix; if it is put after the base, it is a suffix; and if it is inserted within the base, splitting it into two discontinuous parts, it is an infix.) Other morphological processes recognized by linguists need not be

mentioned here, but reference may be made to the fact that many of Bloomfield's followers from the mid-1940s were dissatisfied with the whole notion of morphological processes. Instead of saying that -ing was affixed to sing they preferred to say that sing and -ing co-occurred in a particular pattern or arrangement, thereby avoiding the implication that sing is in some sense prior to or more basic than -ing. The distinction of morpheme and morph (and the notion of allomorphs) was developed in order to make possible the description of the morphology and syntax of a language in terms of "arrangements" of items rather than in terms of "processes" operating upon more basic items. Nowadays, the opposition to "processes" is, except among the stratificationalists, almost extinct. It has proved to be cumbersome, if not impossible, to describe the relationship between certain linguistic forms without deriving one from the other or both from some common underlying form, and most linguists no longer feel that this is in any way reprehensible.

Phonology.

With the great progress made in phonetics in the late 19th century, it had become clear that the question whether two speech sounds were the same or not was more complex than might appear at first sight. Two utterances of what was taken to be the same word might differ quite perceptibly from one occasion of utterance to the next. Some of this variation could be attributed to a difference of dialect or accent and is of no concern here. But even two utterances of the same word by the same speaker might vary from one occasion to the next. Variation of this kind, though it is generally less obvious and would normally pass unnoticed, is often clear enough to the trained phonetician and is measurable instrumentally. It is known that the "same" word is being uttered, even if the physical signal produced is variable, in part, because the different pronunciations of the same word will cluster around some acoustically identifiable norm. But this is not the whole answer, because it is actually impossible to determine norms of pronunciation in purely acoustic terms. Once it has been decided what counts as "sameness" of sound from the linguistic

point of view, the permissible range of variation for particular sounds in particular contexts can be measured, and, within certain limits, the acoustic cues for the identification of utterances as "the same" can be determined.

What is at issue is the difference between phonetic and phonological (or phonemic) identity, and for these purposes it will be sufficient to define phonetic identity in terms solely of acoustic "sameness." Absolute phonetic identity is a theoretical ideal never fully realized. From a purely phonetic point of view, sounds are more or less similar, rather than absolutely the same or absolutely different. Speech sounds considered as units of phonetic analysis in this article are called phones, and, following the normal convention, are represented by enclosing the appropriate alphabetic symbol in square brackets. Thus [p] will refer to a p sound (i.e., what is described more technically as a voiceless, bilabial stop); and [pit] will refer to a complex of three phones - a p sound, followed by an i sound, followed by a t sound. A phonetic transcription may be relatively broad (omitting much of the acoustic detail) or relatively narrow (putting in rather more of the detail), according to the purpose for which it is intended. A very broad transcription will be used in this article except when finer phonetic differences must be shown.

Phonological, or phonemic, identity was referred to above as "sameness of sound from the linguistic point of view." Considered as phonological units - i.e., from the point of view of their function in the language - sounds are described as phonemes and are distinguished from phones by enclosing their appropriate symbol (normally, but not necessarily, an alphabetic one) between two slash marks. Thus /p/ refers to a phoneme that may be realized on different occasions of utterance or in different contexts by a variety of more or less different phones. Phonological identity, unlike phonetic similarity, is absolute: two phonemes are either the same or different, they cannot be more or less similar. For example, the English words "bit" and "pit" differ phonemically in that the first has the phoneme /b/ and the second has the phoneme /p/ in initial position. As the words are normally pronounced, the phonetic realization of /b/ will differ from the phonetic realization of /p/ in a number of different ways: it will be at least partially voiced (i.e., there

will be some vibration of the vocal cords), it will be without aspiration (i.e., there will be no accompanying slight puff of air, as there will be in the case of the phone realizing /p/), and it will be pronounced with less muscular tension. It is possible to vary any one or all of these contributory differences, making the phones in question more or less similar, and it is possible to reduce the phonetic differences to the point that the hearer cannot be certain which word, "bit" or "pit," has been uttered. But it must be either one or the other; there is no word with an initial sound formed in the same manner as /p/ or /b/ that is halfway between the two. This is what is meant by saying that phonemes are absolutely distinct from one another - they are discrete rather than continuously variable.

Semantics.

Bloomfield thought that semantics, or the study of meaning, was the weak point in the scientific investigation of language and would necessarily remain so until the other sciences whose task it was to describe the universe and man's place in it had advanced beyond their present state. In his textbook Language (1933), he had himself adopted a behaviouristic theory of meaning, defining the meaning of a linguistic form as "the situation in which the speaker utters it and the response which it calls forth in the hearer." Furthermore, he subscribed, in principle at least, to a physicalist thesis, according to which all science should be modelled upon the so-called exact sciences and all scientific knowledge should be reducible, ultimately, to statements made about the properties of the physical world. The reason for his pessimism concerning the prospects for the study of meaning was his feeling that it would be a long time before a complete scientific description of the situations in which utterances were produced and the responses they called forth in their hearers would be available. At the time that Bloomfield was writing, physicalism was more widely held than it is today, and it was perhaps reasonable for him to believe that linguistics should eschew mentalism and concentrate upon the directly observable. As a result, for some 30 years after the publication of Bloomfield's textbook, the study of meaning was almost wholly neglected by his

followers; most American linguists who received their training during this period had no knowledge of, still less any interest in, the work being done elsewhere in semantics.

Two groups of scholars may be seen to have constituted an exception to this generalization: anthropologically minded linguists and linguists concerned with Bible translation. Much of the description of the indigenous languages of America has been carried out since the days of Boas and his most notable pupil Sapir by scholars who were equally proficient both in anthropology and in descriptive linguistics; such scholars have frequently added to their grammatical analyses of languages some discussion of the meaning of the grammatical categories and of the correlations between the structure of the vocabularies and the cultures in which the languages operated. It has already been pointed out that Boas and Sapir and, following them, Whorf were attracted by Humboldt's view of the interdependence of language and culture and of language and thought. This view was quite widely held by American anthropological linguists (athough many of them would not go as far as Whorf in asserting the dependence of thought and conceptualization upon language).

Also of considerable importance in the description of the indigenous languages of America has been the work of linguists trained by the American Bible Society and the Summer Institute of

Linguistics, a group of Protestant missionary linguists. Because their principal aim is to produce translations of the Bible, they have necessarily been concerned with meaning as well as with grammar and phonology. This has tempered the otherwise fairly orthodox Bloomfieldian approach characteristic of the group.

The two most important developments evident in recent work in semantics are, first, the application of the structural approach to the study of meaning and, second, a better appreciation of the relationship between grammar and semantics. The second of these developments will be treated in the following section on Transformational-generative grammar.

The first, structural semantics, goes back to the period preceding World War II and is exemplified in a large number of publications, mainly by German scholars - Jost Trier, Leo Weisgerber, and their collaborators.

The structural approach to semantics is best explained by contrasting it with the more traditional "atomistic" approach, according to which the meaning of each word in the language is described, in principle, independently of the meaning of all other words. The structuralist takes the view that the meaning of a word is a function of the relationships it contracts with other words in a particular lexical field, or subsystem, and that it cannot be adequately described except in terms of these relationships. For example, the colour terms in particular languages constitute a lexical field, and the meaning of each term depends upon the place it occupies in the field. Although the denotation of each of the words "green," "blue," and "yellow" in English is somewhat imprecise at the boundaries, the position that each of them occupies relative to the other terms in the system is fixed: "green" is between "blue" and "yellow," so that the phrases "greenish yellow" or "yellowish green" and "bluish green" or "greenish blue" are used to refer to the boundary areas. Knowing the meaning of the word "green" implies knowing what cannot as well as what can be properly described as green (and knowing of the borderline cases that they are borderline cases). Languages differ considerably as to the number of basic colour terms that they recognize, and they draw boundaries within the psychophysical continuum of colour at different places. Blue, green, yellow, and so on do not exist as distinct colours in nature, waiting to be labelled differently, as it were, by different languages; they come into existence, for the speakers of particular languages, by virtue of the fact that those languages impose structure upon the continuum of colour and assign to three of the areas thus recognized the words "blue," "green," "yellow."

The language of any society is an integral part of the culture of that society, and the meanings recognized within the vocabulary of the language are learned by the child as part of the process of acquiring the culture of the society in which he is brought up. Many of the structural differences found in the vocabularies of

different languages are to be accounted for in terms of cultural differences. This is especially clear in the vocabulary of kinship (to which a considerable amount of attention has been given by anthropologists and linguists), but it holds true of many other semantic fields also. A consequence of the structural differences that exist between the vocabularies of different languages is that, in many instances, it is in principle impossible to translate a sentence "literally" from one language to another.

It is important, nevertheless, not to overemphasize the semantic incommensurability of languages. Presumably, there are many physiological and psychological constraints that, in part at least, determine one's perception and categorization of the world. It may be assumed that, when one is learning the denotation of the more basic words in the vocabulary of one's native language, attention is drawn first to what might be called the naturally salient features of the environment and that one is, to this degree at least, predisposed to identify and group objects in one way rather than another. It may also be that human beings are genetically endowed with rather more specific and linguistically relevant principles of categorization. It is possible that, although languages differ in the number of basic colour categories that theydistinguish, there is a limited number of hierarchically ordered basic colour categories from which each language makes its selection and that what counts as a typical instance, or focus, of these universal colour categories is fixed and does not vary from one language to another. If this hypothesis is correct, then it is false to say, as many structural semanticists have said, that languages divide the continuum of colour in a quite arbitrary manner. But the general thesis of structuralism is unaffected, for it still remains true that each language has its own unique semantic structure even though the total structure is, in each case, built upon a substructure of universal distinctions.

Structural linguistics in America.

American and European structuralism shared a number of features. In insisting upon the necessity of treating each language as a more or less coherent and

integrated system, both European and American linguists of this period tended to emphasize, if not to exaggerate, the structural uniqueness of individual languages. There was especially good reason to take this point of view given the conditions in which American linguistics developed from the end of the 19th century. There were hundreds of indigenous American Indian languages that had never been previously described. Many of these were spoken by only a handful of speakers and, if they were not recorded before they became extinct, would be permanently inaccessible. Under these circumstances, such linguists as Franz Boas (died 1942) were less concerned with the construction of a general theory of the structure of human language than they were with prescribing sound methodological principles for the analysis of unfamiliar languages. They were also fearful that the description of these languages would be distorted by analyzing them in terms of categories derived from the analysis of the more familiar Indo-European languages.

After Boas, the two most influential American linguists were Edward Sapir (died 1939) and Leonard Bloomfield (died 1949). Like his teacher Boas, Sapir was equally at home in anthropology and linguistics, the alliance of which disciplines has endured to the present day in many American universities. Boas and Sapir were both attracted by the Humboldtian view of the relationship between language and thought, but it was left to one of Sapir's pupils, Benjamin Lee Whorf, to present it in a sufficiently challenging form to attract widespread scholarly attention. Since the republication of Whorf's more important papers in 1956, the thesis that language determines perception and thought has come to be known as the Whorfian hypothesis.

Sapir's work has always held an attraction for the more anthropologically inclined American linguists. But it was Bloomfield who prepared the way for the later phase of what is now thought of as the most distinctive manifestation of American "structuralism." When he published his first book in 1914, Bloomfield was strongly influenced by Wundt's psychology of language. In 1933, however, he published a drastically revised and expanded version with the new title Language; this book dominated the field for the next 30 years. In it Bloomfield explicitly

adopted a behaviouristic approach to the study of language, eschewing in the name of scientific objectivity all reference to mental or conceptual categories. Of particular consequence was his adoption of the behaviouristic theory of semantics according to which meaning is simply the relationship between a stimulus and a verbal response. Because science was still a long way from being able to give a comprehensive account of most stimuli, no significant or interesting results could be expected from the study of meaning for some considerable time, and it was preferable, as far as possible, to avoid basing the grammatical analysis of a language on semantic considerations. Bloomfield's followers pushed even further the attempt to develop methods of linguistic analysis that were not based on meaning. One of the most characteristic features of "post-Bloomfieldian" American structuralism, then, was its almost complete neglect of semantics.

Another characteristic feature, one that was to be much criticized by Chomsky, was its attempt to formulate a set of "discovery procedures" - procedures that could be applied more or less mechanically to texts and could be guaranteed to yield an appropriate phonological and grammatical description of the language of the texts. Structuralism, in this narrower sense of the term, is represented, with differences of emphasis or detail, in the major American textbooks published during the 1950s.

Structural linguistics in Europe.

Structural linguistics in Europe is generally said to have begun in 1916 with the posthumous publication of the Cours de Linguistique Générale (Course in General Linguistics) of Ferdinand de Saussure. Much of what is now considered as Saussurean can be seen, though less clearly, in the earlier work of Humboldt, and the general structural principles that Saussure was to develop with respect to synchronic linguistics in the Cours had been applied almost 40 years before (1879) by Saussure himself in a reconstruction of the Indo-European vowel system. The full significance of the work was not appreciated at the time. Saussure's structuralism can be summed up in two dichotomies (which jointly cover what Humboldt referred to in terms of his own distinction of inner and outer form): (1)

langue versus parole and (2) form versus substance. By langue, best translated in its technical Saussurean sense as language system, is meant the totality of regularities and patterns of formation that underlie the utterances of a language; by parole, which can be translated as language behaviour, is meant the actual utterances themselves. Just as two performances of a piece of music given by different orchestras on different occasions will differ in a variety of details and yet be identifiable as performances of the same piece, so two utterances may differ in various ways and yet be recognized as instances, in some sense, of the same utterance. What the two musical performances and the two utterances have in common is an identity of form, and this form, or structure, or pattern, is in principle independent of the substance, or "raw material," upon which it is imposed. "Structuralism," in the European sense then, refers to the view that there is an abstract relational structure that underlies and is to be distinguished from actual utterances - a system underlying actual behaviour - and that this is the primary object of study for the linguist.

Two important points arise here: first, that the structural approach is not in principle restricted to synchronic linguistics; second, that the study of meaning, as well as the study of phonology and grammar, can be structural in orientation. In both cases "structuralism" is opposed to "atomism" in the European literature. It was Saussure who drew the terminological distinction between synchronic and diachronic linguistics in the Cours; despite the undoubtedly structural orientation of his own early work in the historical and comparative field, he maintained that, whereas synchronic linguistics should deal with the structure of a language system at a given point in time, diachronic linguistics should be concerned with the historical development of isolated elements - it should be atomistic. Whatever the reasons that led Saussure to take this rather paradoxical view, his teaching on this point was not generally accepted, and scholars soon began to apply structural concepts to the diachronic study of languages. The most important of the various schools of structural linguistics to be found in Europe in the first half of the 20th century have included the Prague school, most notably represented by Nikolay

Sergeyevich Trubetskoy (died 1938) and Roman Jakobson (born 1896), both Russian émigrés, and the Copenhagen (or glossematic) school, centred around Louis Hjelmslev (died 1965). John Rupert Firth (died 1960)
and his followers, sometimes referred to as the London school, were less Saussurean in their approach, but, in a general sense of the term, their approach may also be described appropriately as structural linguistics.

Syntax.

Syntax, for Bloomfield, was the study of free forms that were composed entirely of free forms. Central to his theory of syntax were the notions of form classes and constituent structure. (These notionswere also relevant, though less central, in the theory of morphology.) Bloomfield defined form classes, rather imprecisely, in terms of some common "recognizable phonetic or grammatical feature" shared by all the members. He gave as examples the form class consisting of "personal substantive expressions" in English (defined as "the forms that, when spoken with exclamatory final pitch, are calls for a person's presence or attention" - e.g., "John," "Boy," "Mr. Smith"); the form class consisting of "infinitive expressions" (defined as "forms which, when spoken with exclamatory final pitch, have the meaning of a command" - e.g., "run," "jump," "come here"); the form class of "nominative substantive expressions" (e.g., "John," "the boys"); and so on. It should be clear from these examples that form classes are similar to, though not identical with, the traditional parts of speech and that one and the same form can belong to more than one form class.

What Bloomfield had in mind as the criterion for form class membership (and therefore of syntactic equivalence) may best be expressed in terms of substitutability. Form classes are sets of forms (whether simple or complex, free or bound), any one of which may be substituted for any other in a given construction or set of constructions throughout the sentences of the language.

The smaller forms into which a larger form may be analyzed are its constituents, and the larger form is a construction. For example, the phrase "poor John" is a

construction analyzable into, or composed of, the constituents "poor" and "John." Because there is no intermediate unit of which "poor" and "John" are constituents that is itself a constituent of the construction "poor John," the forms "poor" and "John" may be described not only as constituents but also as immediate constituents of "poor John."

Similarly, the phrase "lost his watch" is composed of three word forms - "lost," "his," and "watch" - all of which may be described as constituents of the construction. Not all of them, however, are its immediate constituents. The forms "his" and "watch" combine to make the intermediate construction "his watch"; it is this intermediate unit that combines with "lost" to form the larger phrase "lost his watch." The immediate constituents of "lost his watch" are "lost" and "his watch"; the immediate constituents of "his watch" are the forms "his" and "watch." By the constituent structure of a phrase or sentence is meant the hierarchical organization of the smallest forms of which it is composed (its ultimate constituents) into layers of successively more inclusive units. Viewed in this way, the sentence "Poor John lost his watch" is more than simply a sequence of five word forms associated with a particular intonation pattern. It is analyzable into the immediate constituents "poor John" and "lost his watch," and each of these phrases is analyzable into its own immediate constituents and so on, until, at the last stage of the analysis, the ultimate constituents of the sentence are reached.

Each form, whether it is simple or composite, belongs to a certain form class. Using arbitrarily selected letters to denote the form classes of English, "poor" may be a member of the form class A, "John" of the class B, "lost" of the class C, "his" of the class D, and "watch" of the class E. Because "poor John" is syntactically equivalent to (i.e., substitutable for) "John," it is to be classified as a member of A. So too, it can be assumed, is "his watch." In the case of "lost his watch" there is a problem. There are very many forms - including "lost," "ate," and "stole" - that can occur, as here, in constructions with a member of B and can also occur alone; for example, "lost" is substitutable for "stole the money," as "stole" is substitutable for either or for "lost his watch." This being so, one might decide to classify

constructions like "lost his watch" as members of C. On the other hand, there are forms that - though they are substitutable for "lost," "ate," "stole," and so on when these forms occur alone - cannot be used in combination with a following member of B (cf. "died," "existed"); and there are forms that, though they may be used in combination with a following member of B, cannot occur alone (cf. "enjoyed"). The question is whether one respects the traditional distinction between transitive and intransitive verb forms. It may be decided, then, that "lost," "stole," "ate" and so forth belong to one class, C (the class to which "enjoyed" belongs), when they occur "transitively" (i.e., with a following member of B as their object) but to a different class, F (the class to which "died" belongs), when they occur "intransitively." Finally, it can be said that the whole sentence "Poor John lost his watch" is a member of the form class G. Thus the constituent structure not only of "Poor John lost his watch" but of a whole set of English sentences can be represented by means of the tree diagram given in Figure 2. New sentences of the same type can be constructed by substituting actual forms for the class labels.

Any construction that belongs to the same form class as at least one of its immediate constituents is described as endocentric; the only endocentric construction in the model sentence above is "poor John." All the other constructions, according to the analysis, are exocentric. This is clear from the fact that in Figure 2 the letters at the nodes above every phrase other than the phrase A + B (i.e., "poor John," "old Harry," and so on) are different from any of the letters at the ends of the lower branches connected directly to these nodes. For example, the phrase D + E (i.e., "his watch," "the money," and so forth) has immediately above it a node labelled B, rather than either D or E. Endocentric constructions fall into two types: subordinating and coordinating. If attention is confined, for simplicity, to constructions composed of no more than two immediate constituents, it can be said that subordinating constructions are those in which only one immediate constituent is of the same form class as the whole construction, whereas coordinating constructions are those in which both constituents are of the same form class as the whole construction. In a subordinating construction (e.g., "poor

John"), the constituent that is syntactically equivalent to the whole construction is described as the head, and its partner is described as the modifier: thus, in "poor John," the form "John" is the head, and "poor" is its modifier. An example of a coordinating construction is "men and women," in which, it may be assumed, the immediate constituents are the word "men" and the word "women," each of which is syntactically equivalent to "men and women." (It is here implied that the conjunction "and" is not a constituent, properly so called, but an element that, like the relative order of the constituents, indicates the nature of the construction involved. Not all linguists have held this view.)

One reason for giving theoretical recognition to the notion of constituent is that it helps to account for the ambiguity of certain constructions. A classic example is the phrase "old men and women," which may be interpreted in two different ways according to whether one associates "old" with "men and women" or just with "men." Under the first of the two interpretations, the immediate constituents are "old" and "men and women"; under the second, they are "old men" and "women." The difference in meaning cannot be attributed to any one of the ultimate constituents but results from a difference in the way in which they are associated with one another. Ambiguity of this kind is referred to as syntactic ambiguity. Not all syntactic ambiguity is satisfactorily accounted for in terms of constituent structure.

TAGMEMICS

The system of tagmemic analysis, as presented by Kenneth L. Pike, was developed for the analysis not only of language but of all of human behaviour that manifests the property of patterning. In the following treatment, only language will be discussed.

Modes of language.

Every language is said to be trimodal - i.e., structured in three modes: phonology, grammar, and lexicon. These modes are interrelated but have a considerable degree

of independence and must be described in their own terms. Phonology and lexicon should not be seen as mere appendages to grammar, the former simply specifying which phonemes can combine to form morphemes (or morphs), and the latter simply listing the morphemes and other meaningful units with a description of their meaning. There are levels of structure in each of the modes, and the units of one level are not necessarily coterminous with those of another. Phonemes, for example, may combine to form syllables and syllables to form phonological words ("phonological word" is defined as the domain of some phonological process such as accentuation, assimilation, or dissimilation), but the morpheme (or morph) will not necessarily consist of an integral number of syllables, still less of a single syllable. Nor will the word as a grammatical unit necessarily coincide with the phonological word. Similarly, the units of lexical analysis, sometimes referred to as lexemes (in one sense of this term), are not necessarily identifiable as single grammatical units, whether as morphemes, words, or phrases. No priority, then, is ascribed to any one of the three modes.

The originality of tagmemic analysis and the application of the term tagmeme is most clearly manifest in the domain of grammar. By a tagmeme is meant an element of a construction, the element in question being regarded as a composite unit, described in such terms as "slot-filler" or "function-class." For example, one of the tagmemes required for the analysis of English at the syntactic level might be noun-as-subject, in which "noun" refers to a class of substitutable, or paradigmatically related, morphemes or words capable of fulfilling a certain grammatical function, and "subject" refers to the function that may be fulfilled by one or more classes of elements. In the tagmeme noun-as-subject - which, using the customary tagmemic symbolism, may be represented as Subject:noun - the subject slot is filled by a noun. When a particular tagmeme is identified in the analysis of an actual utterance, it is said to be manifested by the particular member of the grammatical class that occurs in the appropriate slot in the utterance. For example, in the utterance "John is asleep," the subject tagmeme is manifested by the noun "John." Tagmemicists insist that tagmemes, despite their bipartite

structure, are single units. In grammatical analysis, the distribution of tagmemes, not simply of classes, is stated throughout the sentences of the language. Subject:noun is a different tagmeme from Object:noun, as it is also a different tagmeme from Subject:pronoun.

Список литературы

1. Замяткин Н.Ф. Вас невозможно научить иностранному языку. – Неография, 2006. – 168 с.
2. Ловцевич Г.Н.. Об индивидуальном чтении в старших классах // ИЯШ, 1989. – №6, – С. 28-31.
3. Соловова Е.Н. Методика обучения иностранным языкам. Базовый курс лекций. 3-е издание. – М.: Просвещение, 2006. – 239 с.
4. Цыперсон М.Б. Программа курса «Домашнее чтение на английском языке». – М.: Макмиллан, 2012. – 18 с.
5. Чарекова Е.П., Баграмова Е.В. Практика английского языка (сборник рассказов и упражнений для домашнего чтения). – Ростов на Дону «Феникс», 2004. – 320 с.
6. Hafiz, F.M and Tudor, I. Extensive reading and the development of language skills // ELT Journal, 1989. - № 43 (1). –Pp. 4-13.
7. Hoey, Michael. Lexical Priming: A New Theory of Words and Language. – London: Routledge, 2005. – 202 p.
8. Kroll, Barbara. Exploring the Dynamics of Second Language Writing.: Chapter 10 Reading and Writing Relations. – New York: Cambridge University Press, 2003. 15 p.
9. http://www.teachingenglish.org.uk/articles/extensive-reading