Psycholinguistics is the study of the mental processes and skills underlying the production and comprehension of language, and of the acquisition of these skills. This chapter will deal with the former aspect only; for the acquisition of language see the suggested “Further reading” at the end of this chapter.

Although the term “psycholinguistics” was brought into vogue during the 1950s, the psychological study of language use is as old as psychology itself. As early as 1879, for instance, Francis Galton published the first study of word associations (Galton, 1879). And the year 1900 saw the appearance of Wilhelm Wundt’s monumental two-volume work *Die Sprache*. It endeavoured to explain the phylogeny of language in the human mind as an increasingly complex and conscious means of expression in a society, and to describe how language is created time and again in the individual act of speaking. Although Wundt deemed it impossible to study language use experimentally, his contemporaries introduced the experimental study of reading (Huey), of
verbal memory and word association (Ebbinghaus, Marbe, Watt), and of sentence production (Bühler, Seltz). They began measuring vocabulary size (Binet), and started collecting and analysing speech errors (Meringer and Mayer). The study of neurologically induced language impairments acquired particular momentum after Paul Broca and Carl Wernicke discovered the main speech and language supporting areas in the brain's left hemisphere. In the absence of live brain tomography, aphasiologists began developing neurolinguistic tests for the purpose of localizing brain dysfunctions.

All of these themes persist in modern psycholinguistics. But developments since the 1950s have provided it with two of its most characteristic features, which concern linguistic processing and representation. With respect to processing, psycholinguistics has followed mainstream psychology in that it considers the language user as a complex information processing system. With respect to representation, psycholinguists stress the gigantic amount of linguistic knowledge the language user brings to bear in producing and understanding language. Although the structure of this knowledge is the subject matter of linguistics, it is no less a psychological entity than is language processing itself (Chomsky, 1968). Psycholinguistics studies how linguistic knowledge is exploited in language use, how representations for the form and meaning of words, sentences, and texts are constructed or manipulated by the language user, and how the child acquires such linguistic representations.

I shall first introduce the canonical setting for language use: conversation. Next I shall consider the mental lexicon, the heart of our linguistic knowledge. I shall then move to the processes of speaking and speech understanding respectively. Finally I shall turn to other modes of language use, in particular written language and sign language.

CONVERSATION

Our linguistic skills are primarily tuned to the proper conduct of conversation. The innate ability to converse has provided our species with a capacity to share moods, attitudes, and information of almost any kind, to assemble knowledge and skills, to plan coordinated action, to educate its offspring, in short, to create and transmit culture. And all this at a scale that is absolutely unmatched in the animal kingdom. In addition, we converse with ourselves, a kind of autostimulation that makes us more aware of our inclinations, of what we think or intend (Dennett, 1991). Fry (1977) correctly characterized our species as homo loquens.

In conversation the interlocutors are involved in negotiating meaning. When we talk, we usually have some kind of communicative intention, and the conversation is felicitous when that intention is recognized by our partner(s) in conversation (Grice, 1968; Sperber & Wilson, 1986). This may take several turns of mutual clarification. Here is an example from Clark and
Wilkes-Gibbs (1986), where subjects had to refer to complex tangram figures:

A: Uh, person putting a shoe on.
B: Putting a shoe on?
A: Uh huh. Facing left. Looks like he's sitting down.
B: Okay.

Here the communicative intention was to establish reference, and that is often a constituting component of a larger communicative goal. Such goals can be to commit the interlocutor or oneself to some course of action, as in requesting and promising, or to inform the interlocutor on some state of affairs, as in asserting, for example. The appropriate linguistic acts for achieving such goals are called *speech acts* (Austin, 1962).

Although what is said is the means of making the communicative intention recognizable, the relation between the two can be highly indirect. Conversations involve intricate mechanisms of politeness control (Brown & Levinson, 1987). What is *conveyed* is often quite different from what is *said*. In most circumstances, for instance, we don’t request by commanding, like in “Open the window”. Rather we do it indirectly by checking whether the interlocutor is able or willing to open the window, like in “Can you open the window for me?” It would, then, be inappropriate for the interlocutor to answer “Yes” without further action. In that case, the response is only to the question (whether he or she is able to open the window), but not to the request.

How does the listener know that there is a request in addition to the question? There is, of course, an enormous amount of shared situational knowledge that will do the work. Grice (1975) has argued that conversations are governed by principles of rationality; Sperber and Wilson (1986) call it the *principle of relevance*. The interlocutor, for instance, is so obviously able to open the window that the speaker’s intention cannot have been to check that ability. But Clark (1979) found that linguistic factors play a role as well. If the question is phrased idiomatically, involving *can* and *please*, subjects interpret it as a request. But the less idiomatic it is (like in “Are you able to...”), the more subjects react to the question instead of to the request.

Another important aspect of conversation is *turn-taking*. There are rules for the allocation of turns in conversation that ensure everybody’s right to talk, that prevent the simultaneous talk of different parties, and that regulate the proper engaging in and disengaging from conversation (Sacks, Schegloff, & Jefferson, 1974). These rules are mostly followed, and sometimes intentionally violated (as in interrupting the speaker). Turn-taking is subtly controlled by linguistic (especially prosodic) and non-verbal (gaze and body movement) cues (Beattie, 1983).

THE MENTAL LEXICON

Producing or understanding spoken language always involves the use of
words. The mental lexicon is our repository of words, their meanings, their syntax, and their sound forms. A language’s vocabulary is, in principle, unlimited in size. Take, for instance, the numerals in English. They alone form an infinite set of words. But it is unlikely that a word such as *twenty-three-thousand-two-hundred-and-seventy-nine* is an entry in our mental lexicon. Rather, such a word is constructed by rule when needed. We have the ability to produce new words that are not stored in our mental lexicon.

![Diagram of a lexical network](image)

*Figure 1* Fragment of a lexical network. Each word is represented at the conceptual, the syntactic and the sound form level.

*Source*: Bock and Levelt, 1993
How many words are stored? Miller (1991) estimates that the average high school graduate knows about 60,000 words (under one definition of “word”).

One way of representing this enormous body of knowledge is by way of network models. Figure 1 shows a fragment of such a network. Each word is represented by three nodes, one at the conceptual level, one at the syntactic (grammatical) or lemma level, and one at the sound form (phonological) or lexeme level. The lemma is the syntactic representation and the lexeme is the phonological representation. A word’s semantic properties are given by its connections to other nodes at the conceptual level (for instance, that a sheep is an animal, gives milk, etc.). A word’s syntactic properties are represented by its lemma node’s relations to other syntactic nodes (for instance, “sheep” is a noun; French “mouton” has male gender, etc.). The sound form properties, finally, such as a word’s phonological segments, are represented in the way a word’s lexeme node relates to other sound form nodes (“sheep” for instance contains three ordered phonological segments, /ʃ/, /i/, and /p/, as shown in Figure 1).

Different authors have proposed different network models (e.g., Collins & Loftus, 1975; Dell, 1986; Roelofs, 1992), and for different purposes. It is unlikely that such networks can adequately represent all complexities of our semantic, syntactic, and phonological knowledge about words. But they can be useful in predicting speed of word access in comprehension and production, as well as in explaining various kinds of errors that we make in speech production and various disorders of accessing words in aphasic speech.

Especially important for theories of language use are the ways that verbs are represented in the mental lexicon. As a semantic entity, a verb assigns semantic roles to its arguments. The verb walk, for instance, requires an animate argument that specifies the role of agent, as in John walked. The verb greet governs two arguments, one for the agent and one for the recipient of the action, as in Peter greeted the driver. As a syntactic entity, a verb assigns syntactic functions to the sentence constituents it governs. In the above sentence, Peter is the subject and the driver the object. A verb’s argument-function mapping is not random. Most verbs, for instance, map a recipient argument on to a syntactic object function, but not all. The verb receive doesn’t. In Mary received the book, Mary is both recipient and sentence subject. Also, verbs often allow for multiple mappings. In the driver was greeted by Peter, the recipient, not the agent appears in subject position.

For each verb, the mental lexicon contains its possible mapping frames. These play an important role in the speaker’s syntactic planning and in the listener’s syntactic and semantic parsing.

SPEAKING

Speaking is our most complex cognitive-motor skill. It involves the conception of an intention, the selection of information whose expression will make
that intention recognizable, the selection of appropriate words, the construction of a syntactic framework, the retrieval of the words’ sound forms, and the computation of an articulatory plan for each word and for the utterance as a whole. It also involves the execution of this plan by more than 100 muscles controlling the flow of air through the vocal tract. Finally, it involves a process of self-monitoring by which speech trouble can be prevented or repaired. The following is a bird eye’s view over these processes.

**Conceptual preparation**

The question where communicative intentions come from is a psychodynamic question rather than a psycholinguistic one. Speaking is a form of social action, and it is in the context of action that intentions, goals, and subgoals develop. It is not impossible, though, that the intention what to say occasionally arises from spontaneous activity in the speech formulating system itself. It can create rather incoherent “internal speech”, which we can self-perceive. This, in turn, may provide us with tatters of notions that we then consider for expression (cf. Dennett, 1991).

Conveying an intention may involve several steps or “speech acts”. The speaker will have to decide what to express first, what next, and so on. This is called the speaker’s linearization problem (Levelt, 1989). It is especially apparent in the expression of multidimensional information, as in describing one’s apartment (Linde & Labov, 1975). The conceptual preparation of speech, and in particular linearization, require the speaker’s continuing attention. The principles of linearization are such that attentional load is minimized.

Each speech act, be it a request to do $X$, an assertion that $Y$, etc., involves the expression of some conceptual structure, technically called a “message” (Garrett, 1975). That message is to be given linguistic shape; it has to become “formulated”.

**Grammatical encoding**

A first step in formulating is to retrieve the appropriate words from the mental lexicon and to embed them in the developing syntactic structure. In normal conversation we produce some two words per second. At this rate we manage to access the appropriate words in our huge mental lexicon. Occasional errors of lexical selection (such as “Don’t burn your toes” where fingers was intended) show that the lexicon has a semantic organization.

The standard explanation for such errors is that activation spreads through a semantically organized network, as in Figure 1. In such a network, each node has an activation level between 0 and 1. When the lexical concept node SHEEP is active, then activation spreads to semantically related concept nodes, such as GOAT. Both nodes spread activation “down” to their lemma.
nodes. Which one of the lemmas will then be selected for further processing? Normally it will be the most activated one, in this case the lemma for "sheep". But the occurrence of an occasional error shows that there is a small probability that a less activated lemma gets selected. According to one theory (Roelofs, 1992) the probability that a particular lemma becomes selected within a time interval $t$ is the ratio of its activation to the sum of the activation of all other lemma nodes. For instance, if "sheep" and "goat" are the only two active lemmas during interval $t$ after presentation of the picture, and they have activation levels of 0.7 and 0.1 respectively, the probability that the target word "sheep" will be selected during that interval is 7/8, whereas the erroneous word "goat" will be selected with the probability 1/8. Hence, if there is more than one lemma active in the system, there is always a small probability that a non-intended word becomes selected (and it is likely to be semantically related to the target).

Spreading activation theories of lexical selection are typically tested in picture-naming experiments, where naming latencies are measured. For a review of issues in lexical selection, see Levelt (1992a).

As soon as a lemma is retrieved, its syntactic properties become available. Among them are the lemma’s grammatical class (preposition, noun, verb, etc.). Each lemma requires its own specific syntactic environment or "frame". Syntactic planning is like solving a set of simultaneous equations. Each lemma’s frame has to fit its neighbour’s frames, and since Garrett (1975) there are theories about how this is realized (see Levelt, 1989, for a review). Actually, the equations are not quite “simultaneous”; the lemmas for an utterance are typically not concurrently retrieved. Lemmas for salient concepts, such as animate objects, tend to be retrieved faster than for non-salient concepts (Bock & Warren, 1985), and that affects their position in the developing syntactic structure. For a review of grammatical encoding, see Bock and Levelt (1994).

**Phonological encoding**

A selected lemma (but only a selected one: see Levelt et al., 1991) spreads its activation to its lexeme node (cf. Figure 1). At this level two kinds of phonological information become available. The first one is the word’s segments, which are “spelled out” one after another. The second one is the word’s metrical structure. For “sheep” it is the information that it is a one-syllable word. For “father” it is the information that it is a two-syllabic trochaic word. The metrical frames of successive words are often combined, creating so-called phonological word frames. In *Peter gave him it*, the last three words form one phonological word *gavimit*. In a process of *segment-to-frame association* spelled-out segments are inserted one by one into the corresponding phonological word frames. It is during this ordered insertion that phonological syllables are created, one after another (such as *ga-vi-mit*; see
Levelt, 1992b). How this string of phonological syllables determines the precise articulatory gestures to be made by the speech organs is still a matter of much debate (see especially Browman & Goldstein, 1991).

The notion that segments and frames are independently retrieved arose in the analysis of phonological speech errors (Dell, 1986; Shattuck-Hufnagel, 1979). Spoonerisms such as *with this wing I thee red*, or *fool the pill* (instead of *fill the pool*) show that segments can become associated to the right place in the wrong frame.

Phonological encoding also involves the planning of larger units than phonological words. There is, in particular, the planning of intonational phrases. These are units that carry a particular intonational contour. Such contours can be rising, falling or combinations thereof. They often express a speaker’s attitude towards what is said: doubt, certainty, or towards the interlocutor: reassuringness, inviting reaction. See Levelt (1989) for a review of phonological encoding.

The output of phonological encoding is an articulatory programme. Phenomenologically, it appears to the speaker as internal speech. This internal speech need not be articulated. It can be kept in an articulatory buffer, ready to be retrieved for articulatory execution (Sternberg, Wright, Knoll, & Monsell, 1980).

**Articulation**

The articulatory apparatus consists of three major structures. The respiratory system controls the steady outflow of air from the lungs. The breathing cycle during speech is quite different from normal breathing, with very rapid inhalation and very slow exhalation. The laryngeal system has the vocal cords as its central part. It is the main source of acoustic energy. The vocal tract, finally, contains the cavities of pharynx, mouth, and nose. They are the resonators that filter the acoustic energy in frequency bands or *formants*. Vowels are characterized by their formant structure. The vocal tract can be constricted at different places, and these constrictions can be made or released in different manners. In this way a wide range of consonantal and other speech sounds can be made.

The control of this utterly complex motor system has been the subject of much research. Present theories converge on the notion of *model-referenced control* (Arbib, 1981; see also Figure 2). The motor system is given an “articulatory task” (as part of the articulatory programme), such as “close the lips”. There are usually many degrees of freedom in executing such a task. For instance, lip closing can be realized by moving the lips, by moving the jaw, or by doing both to various degrees. The internal model computes the least energy-consuming way of reaching the goal, given the actual state of the articulators (there is continuous proprioceptive feedback to the internal model). The output is a set of efferent control signals to the relevant

**Self-monitoring**

We can listen to our own overt speech and detect trouble, just as we can listen to the speech of others and detect errors or infelicitous delivery. This involves our normal speech understanding system. We can also detect trouble in our internal speech. When the trouble is disruptive enough for the ongoing conversation, a speaker may decide to interrupt the flow of speech and to make a self-repair.

Not all self-produced trouble (such as errors of selection) is detected by the speaker. Self-monitoring requires attention; we mostly attend to what we say, far less to how we do it. Detection of trouble is better towards the end of clauses, where less attention for content is required (Levelt, 1989). There are two main classes of trouble that induce repairing. The first one is an all-out error (as in *and above that a horizon*, *no a vertical line*); the error can be lexical, syntactic, or phonological. The second one is that something is not really appropriate (as in *to the right is blue – is a blue point*). The speaker then repairs in order to make the utterance more precise, less ambiguous. Upon detecting either kind of trouble, the speaker can self-interrupt. And this ignores linguistic structure; a speaker can stop in the midst of a phrase, a word, or a syllable. But then, the speaker often marks the kind of trouble
by some editing expression: “no”, “sorry”, “I mean”, for errors; “rather”, “that is”, for something inappropriate.

Restarting, that is, making the repair proper, is linguistically quite principled. The speaker grafts the repair on to the syntax of the interrupted utterance, which has been kept in abeyance. As a consequence, repairing is like linguistic coordination. One seldom finds a repair such as is she driving – she walking downtown? And indeed, the corresponding coordination is she driving or she walking downtown? is ill-formed. But is he – she walking downtown? is a very common repair type, and it corresponds to a well-formed coordination: is he or she walking downtown? (Levelt, 1989).

SPEECH UNDERSTANDING

The canonical objective in speech understanding is to recognize the speaker’s communicative intention. How does the listener induce that intention from the speaker’s overt speech, a continuous flow of acoustic events?

Several component processes are involved here. First, there is the hearer’s acoustic-phonetic analysis of the speech signal, that is, representing it as a phonetic not just an acoustic event. Second, there is phonological decoding, in particular finding the words that correspond to the phonetic events, and analysing the overall prosodic structure of the utterance. Third, there is grammatical decoding, parsing the utterance as a meaningful syntactic structure. Finally, there is discourse processing, interpreting the utterance in the context of the ongoing discourse, and in particular inferring the speaker’s intentions. Let us review these processes in turn.

Acoustic-phonetic analysis

It is very hard, if not impossible, to listen to speech as if it were just a string of chirps, buzzes, hums, and claps. We just cannot help perceiving it as speech. In this so-called “speech mode” (Liberman & Mattingly, 1985) we interpret the acoustic event as resulting from a speaker’s articulatory gestures as a phonetic event. There is no unanimity in the literature, though, about what kind of representation the listener derives. According to Liberman and Mattingly, the listener derives the speaker’s intended articulatory gestures (even if they were sloppy). Others argue that listeners have special detectors for distinctive events in the speech signal, such as for onsets, for spectral peaks, for the frequencies and motions of formants. The detection of such acoustic events may suffice to derive the presence or absence of phonetic features, such as voicing, nasality, vowel height, stridency, and so on (Stevens & Blumstein, 1981).

Speech segments, clusters, and syllables have characteristic distributions of phonetic features. Hence, if such feature detectors are reliable, they may provide sufficient information for effective phonological decoding. Opinions
differ, however, about their reliability. The speech signal is highly variable, dependent as it is on speech rate, sex of the speaker, sloppiness of speech delivery, reverberation or noise in the room, for example. Even if the listener can partial out such effects of the speech context, acoustic-phonetic analysis will often be indeterminate. Still, it may well be sufficient for the purpose. Not every word has to be recognized in order to derive the speaker’s intentions. And where a really critical word is missed, the interlocutor will say “what?” or signal difficulty of understanding in other ways.

For an excellent review of acoustic-phonetic processing, see Pisoni and Luce (1987).

**Phonological decoding**

Whatever the precise character of the phonetic representations, they are the listener’s access codes to the mental lexicon. How does a listener recognize words in connected speech? A major problem here is to segment the speech, to find out where words begin and end in the continuous flow of speech. There are, basically, two routes here.

The first one is the bottom-up approach, that is, to build on cues in the phonetic representation. Cutler (1990) has argued that English listeners will, by default, segment speech such that there are word boundaries right before stressed syllables. It is a statistical fact of English that 85 per cent of the meaningful words that one encounters while listening begin with a stressed syllable. The segmentation strategy will, therefore, be quite successful. Cutler’s theory has meanwhile found substantial experimental support. Also, there are speech sounds that tend to occur at the ends of words, such as [-ng] and [-nd] for English. Speakers may use such phonotactic properties of their language to predict word boundaries.

The second route is top-down. We often recognize a word before it ends. But that means that we can predict the word’s end, and hence the upcoming word boundary. That gives us a handle on where to start recognizing the subsequent word.

Given that we know a word’s beginning, how do we recognize it? According to the cohort theory (Marslen-Wilson, 1989), a small word-initial feature pattern (corresponding to about two segments of the input word) activates all words in the mental lexicon that match it phonologically. Assume the input word is trespass, and the cluster [tr] has become available. This will activate all words beginning with [tr], such as tremble, trespass, trestle, trombone, etc. This is called the “word-initial cohort”. As more phonetic information becomes available, the cohort is successively reduced. When the vowel [e] is perceived, all items not sharing that vowel, such as trombone, are deactivated. This process continues until a single candidate remains. For trespass this happens when [p] is reached. The segment [p] is, therefore, called the uniqueness point of trespass. A word’s uniqueness point depends
on its word-initial lexical alternatives. For most words the uniqueness point precedes the word’s end.

For an optimally efficient system, the word’s uniqueness point would also be its recognition point. There is good experimental evidence in support of this hypothesis (e.g., Frauenfelder, Segui, & Dijkstra, 1990), though the recognition point may slightly anticipate the uniqueness point in case syntactic or semantic information disambiguates the item from its remaining alternatives (Zwitserlood, 1989). Hence, it will often be possible for a listener to anticipate the upcoming word boundary.

Phonological decoding serves not only the recognition of words, but also their groupings into prosodic constituents, such as phonological and intonational phrases. These constituents carry important information about the syntax of the utterance, and about the communicative intentions of the speaker (cf. Levelt, 1989).

**Grammatical decoding**

As words are successively recognized and prosodically grouped, the listener will as much as possible interpret these materials “on-line” (Marslen-Wilson & Tyler, 1980). Each recognized word makes available its syntactic and semantic properties. There is, then, concurrent syntactic parsing and semantic interpretation, each following its own principles, but interacting where necessary.

In this connection, one should distinguish between local and global syntactic parsing. Local parsing involves the creation of local phrase structure, combining words into noun phrases, verb phrases, etc. There is increasing evidence that local parsing can run on word category information alone (Frazier, 1989; Tyler & Warren, 1987). We have little trouble parsing “jabberwocky” or semantically anomalous prose such as *the beer slept the slow guitar*. Here we construct phrase structure exclusively by recognizing the words’ syntactic categories (Art, Adj, N, V). However, successful local parsing is highly dependent on the intactness of phonological phrases, as Tyler and Warren (1987) could show. For instance, in the above anomalous prose, one should not create a prosodic break between *the* and *slow*, or between *slow* and *guitar*.

Global syntactic parsing, however, interacts with semantic interpretation. In global parsing, semantic roles are assigned to syntactic constituents, and this is to a large extent governed by the verb’s argument/function mapping. When the meaning of words or phrases contradicts the semantic roles they should carry, global parsing is hampered (Tyler & Warren, 1987).

One important aspect of global parsing is the resolution of anaphora. In the sentence *the boxer told the skier that the doctor for the team would blame him for the recent injury*, the anaphor *him* can refer back to *the boxer* and to *the skier*, but global syntax prohibits its referring to *the doctor*. Indeed,
experimental evidence shows reactivation of both boxer and skier, but not of doctor when the pronoun him is perceived. Such reactivation can also be measured for so-called null-anaphors as in the policeman saw the boy that the crowd at the party accused t of the crime. Here there is measurable reactivation of boy at position t (the syntactic “trace” of the boy; see Nicol & Swinney, 1989). But also in this respect global parsing is semantically facilitated, for instance if the anaphor’s referent is a concrete noun (Cloitre & Bever, 1988).

Grammatical decoding doesn’t remove all ambiguity (for instance, the pronoun him above is not fully resolved). Here, further discourse processing is needed.

**Discourse processing**

Partners in conversation construct mental models of the state of affairs they are talking about (Johnson-Laird, 1983; Seuren, 1985). Indefinite expressions (such as in there is a dog in the room) make them introduce a new entity (a dog) in the model. Definite expressions (such as the room in the same sentence) make them look up an already existing entity. The new information in the utterance is then attached to whichever entity it concerns.

Identifying referents is a major accomplishment of human language processing, still unmatched by any computer program. The problem is that referring expressions can be highly indirect. How can a waitress in a restaurant interpret the referent when her colleague says the hamburger wants the bill? Nunberg (1979) argued that there are “referring functions” that map a demonstratum (like the hamburger) on to the intended referent (the person who ordered it). But the range of possible referring functions is almost unlimited. Clark, Schreuder, and Buttrick (1983) and Morrow (1986) have argued (and experimentally shown) that such demonstratum-to-referent mapping depends on the mutual knowledge of the interlocutors and on the saliency of entities in their discourse models.

Indirectness is the hallmark of discourse interpretation. As mentioned above, what is said often relates quite indirectly to what the speaker intends to convey. It is not only politeness that governs such indirectness. All figures of speech, whether polite or not, require the listener to build a bridge from the literal to the intended. This holds equally for metaphor (Sperber & Wilson, 1986), irony (Clark & Gerrig, 1984), and hyperbole (Grice, 1975).

Finally, whereas acoustic-phonetic, phonological, and grammatical decoding are largely automatic processes, discourse processing requires the listener’s full attention. In that respect, it is on a par with the speaker’s conceptual preparation. As interlocutors we are concerned with content. The processing of form largely takes care of itself.
The invention of writing systems, whether logographic, syllabic, or alphabetic, is probably the most revolutionary step in human cultural evolution. It added a powerful means of storing and transmitting information. With the invention of printing, it became a major mechanism for large-scale dissemination of knowledge in a culture.

But equally surprising as this ability to map spoken language on to a visual code is our capacity to efficiently process such a code. When skilled, we silently read five or six printed words per second; this is about twice the rate of conversational speech. This ability has not given us any selective advantage in biological evolution; the invention of writing systems is as recent as about 5,000 years ago. Rather, the ability to read must be due to a happy coincidence of other pre-existing faculties of mind.

One of these is, of course, language. As readers we largely use our parsing potential for spoken language. Visual word recognition feeds into the lemma level of Figure 1. As lemmas are successively activated by the printed words, further syntactic, semantic, and discourse processing operates roughly as for spoken language. There are, admittedly, differences too. There is, for instance, no prosody to help syntactic parsing; instead there is punctuation. Also, there is no external enforcement of rate as there is in speech perception.

Another pre-existing faculty on which reading is parasitic is our enormous ability to scan for small meaningful visual patterns. In a hunter’s society these were probably animal silhouettes, footprints, and so on. Words (if not too long or too infrequent) are recognized as wholes; a skilled reader processes a word’s letters in parallel. Much ink has been spilled on the question whether the letters individually or the word as a whole activate a phonological code in silent reading, that is, the word’s lexeme (see Figure 1). Such phonological recoding indeed exists. But it is only for low-frequent words that this “phonological route” is of any help in lemma access (Jared & Seidenberg, 1991). However, this silent “internal speech” probably does play a role in further syntactic and semantic parsing; it is a way of buffering successive words for further processing.

The ability to scan is optimally used in reading. The basic cycle is this: the reader fixates a word for, on average, one-fifth of a second. The fixation is roughly between the beginning and the middle of the word. During this period lexical access is achieved. In addition, there is some perception of the next word in the periphery of vision. Sometimes this suffices to recognize that next word as well on the same fixation (but the fixation will then last somewhat longer). Usually, however, the information from the periphery of vision is used only to plan a saccadic eye movement (a jump of the eye) to that next word. The size of the saccade depends on the length of the next word; the average saccade is about eight characters in size. The new word is fixated, and the cycle starts all over again.
When a word is quite infrequent, or when the reader has trouble integrating it in the developing syntax or semantics, the fixation duration can be substantially longer. Also, the reader may backtrack and refixate an earlier word when there is serious trouble in comprehension.

For a major review of the reading process and its disorders, see Rayner and Pollatsek (1989).

SIGN LANGUAGE

Contrary to written language, the sign languages of deaf people are not parasitic on spoken language. They are autonomous languages in the visual mode. Their mere existence shows that our faculty of language is not crucially

---

**Figure 3** Minimal contrasts between signs in American Sign Language: (a) hand configuration, (b) place of articulation, (c) movement

*Source: From Klima and Bellugi, 1979*
dependent on our ability to speak. Deaf children who grow up in a signing deaf community acquire their language at the same age and in roughly the same stages as hearing children do.

Just as words, signs have form and meaning. The articulators of sign language are the hands, the face, and the body. Where words contrast phonemically (for instance in voicing: bath vs path), signs contrast in hand configuration, in place of articulation and in hand movement (see Figure 3). Also, facial features may distinguish between signs.

Although the first coining of a sign is often iconic, its meaning is eventually independent of its form, as it is for words in spoken languages. As a consequence, sign languages are mutually unintelligible, just as spoken languages are (contrary to what Wundt suggested in Die Sprache – see above).

Sign languages are rich in morphology (for inflection and for derivation of new signs) and have full-fledged recursive syntax. Many syntactic devices are spatial in character. Anaphora, that is, referring back to an earlier introduced entity, is done by pointing to the locus in the signing space (in front of the body) where the original referent was first “established”. In American Sign Language the sign for transitive verbs either moves from subject to object locus, or from object to subject locus. Each verb has its own “mapping function” (like in spoken language, see above). For the structure and use of British Sign Language, see Kyle and Woll (1985).

There is increasing evidence that a sign language is subserved by the same areas of the brain that sustain spoken language. Poizner, Klima, and Bellugi, (1987) showed that damage to anterior areas of the left hemisphere in native signers resulted in a style of signing highly comparable to the agrammatism of so-called Broca’s patients. Similarly, a form of fluent aphasia resulted when the damage was in a more posterior area of the left hemisphere, comparable to the fluent aphasia of so-called Wernicke’s patients. Damage in the right hemisphere left the signing intact, but patients lost the ability to sign coherently about spatial relations, such as the layout of their apartment. Their spatial representations were damaged, but not their spatial language.

**FURTHER READING**

REFERENCES


The nature of human thought and the capacity for rational reasoning have been issues of great interest to philosophers and psychologists since the time of Aristotle. Humans have excelled among species in their ability to solve problems and to adapt their environment for their own purposes. We are unique in our possession of a highly sophisticated system of language allowing both representation of complex and abstract concepts and the communication of very precise meaning with one another. We have also developed a new form of evolution — much faster than natural selection — whereby the accumulated knowledge and wisdom of our culture is recorded and passed on through education so that each new generation starts with an advantage on the one before. Despite this impressive record, we also are subject to many systematic errors and biases in our thinking, some of which are discussed in this chapter.

The study of thinking and reasoning in humans can accurately be described as the study of the nature of intelligence. The work described here falls, however, into a quite different tradition from the psychometric study of individual differences in intelligent performance that is usually referred to as the psychology of intelligence. Psychometrics is concerned with the measurement of intelligent performance, whereas the study of thinking and reasoning is
THINKING AND REASONING

focused on understanding the nature of intelligent processes. Strangely enough, these turn out to be two quite different kinds of undertaking.

THE NATURE OF THINKING: AN HISTORICAL PERSPECTIVE

Historically, we can trace three different conceptions of the nature of thinking. The first of these corresponds to what the non-psychologist might respond if asked to define thought. I shall describe this notion as the contents of consciousness. Common sense (or folk psychology) supposes that we are consciously in control of our actions: we think, therefore we do. When we make a decision or solve a problem it is on the basis of a train of thought of which we are conscious and which we can, if required, describe to another. Such reports of thought are known as introspections. The validity of introspection is clearly assumed in our everyday folk psychology, as we all feel able to ask and answer questions about how and why we have taken particular actions. Indeed, a major industry – opinion polling – is based upon introspectionism. Politicians and political commentators alike are absorbed by the results of polls that ask people not only how they intend to vote, but also to identify the issues which will influence their decisions.

Aristotle and other early philosophers were in no doubt that the mind could and should study itself through introspection. This led to a theory of thinking known as associationism in which thinking was supposed to consist of a sequence of images linked by one of several principles (see Mandler & Mandler, 1964). Associationism and the equation of thought with consciousness remained more or less unchallenged until the late nineteenth and early twentieth centuries when several separate developments conspired to challenge this idea.

First, there were the systematic experimental studies of introspection carried out at the Würzburg School around the beginning of the twentieth century (see Humphrey, 1951). In these experiments, subjects were asked to perform simple cognitive acts such as giving word associations or judging the comparative weight of two objects and then asked to report on what went through their minds at the time. Much to the initial surprise of the researchers, many of these acts did not appear to be mediated by conscious thoughts. Subjects often reported either no conscious experience at all, or else one of indescribable or “imageless” thought.

A second influential development was that of the Freudian school of psychoanalysis which introduced the notion of unconscious thought and motivation. An introspective report of the reason for an action would certainly be suspect to a Freudian since it might well constitute a rationalization of behaviour determined by deep-seated and repressed emotions in the unconscious mind.

The other major influence was the introduction of the school of behaviourism by J. B. Watson (e.g., 1920) whose influence was very strong in
psychology up until the 1950s and which lingers on even in the present day. Watson attacked all study of conscious thought as mentalistic and unscientific. Science, he maintained, could concern itself only with the study of phenomena that were subject to objective observation and independent verification — criteria that introspective reports clearly could not meet. Watson and other behaviourists effectively redefined thought as simply complex forms of behaviour which were the result of stimulus–response learning. Study of stimulus–response pairings and reinforcement history were sufficient to explain all phenomena attributed — by the mentalistically inclined — to thinking.

From the viewpoint of a modern cognitive psychologist both introspectionists and behaviourists might be seen as half right. The behaviourists were probably right in their contention that thought cannot be studied effectively via introspection. The mentalists, on the other hand, were correct in asserting that complex behaviour could not be explained without reference to internal mental processes. Their mistake — with the benefit of hindsight — was to assume that such processes were necessarily conscious and reportable. This leads us to the third conception of human thought — that of information processing.

Psychologists' own thinking — like that of their subjects — is constrained by the availability of models and analogies. Watson used the analogy of a telephone exchange to explain his notion of learning by stimulus–response connections. Although its origin can be traced to earlier, highly creative thinkers (especially Craik, 1943) the emergence of cognitive psychology in the 1950s and 1960s was largely due to the development of cybernetic systems and then the digital computer. Computers are general-purpose information processing systems. They compute by manipulating symbols which can represent almost anything — numbers and arithmetical operators, permitting arithmetic; letters and words as in word and text processing; collections of facts stored in a database; and so on. When people perform mental arithmetic, we would describe this as an act of thought. So is a computer also thinking when it performs computations to solve problems? It appears that it is, although some philosophers (e.g., Searle, 1980) maintain that computer intelligence is intrinsically different from that of the human mind. The point of the analogy, however, is that we can see that computers can perform complex acts of information processing — depending upon their programming — but without any need to assume that they are conscious. Once you equate thinking with information processing, then the task of the modern cognitive psychologist is clear: understanding thought is the problem of discovering the software of the human brain. Many psychological theories in fact are formulated as working computer programs which attempt to simulate the behaviour of a human being who is solving a problem or engaged in some other cognitive activity.

In spite of this advance, arguments persist among cognitive and social
psychologists as to the value of introspective reports. Some cognitive psychologists disregard them entirely on the basis of much evidence that such reports can be both incomplete and misleading (Nisbett & Wilson, 1977). One interesting line of argument is that verbal reports are useful indicators of thought processes but not as used in the tradition of introspective reporting (Ericsson & Simon, 1980). According to this view, verbalizations are the products of cognitive processes and can be fruitfully interpreted by the psychologist when subjects are asked to “think aloud” while performing a task or solving a problem. Introspective reports fail because first, they are retrospective rather than concurrent, and second, they invite subjects to describe their thinking or to theorize about the causes of their behaviour.

The psychology of thinking can be broadly defined to cover a wide range of topics. For example, Gilhooly (1982) distinguishes between directed thinking – as found in problem solving and reasoning – undirected thinking – as in day-dreaming – and creative thinking. In this chapter we shall focus on directed thinking: thought aimed at achieving specific goals. This is an area in which reasonable theoretical progress has been made, and for which there are clear practical applications in everyday life.

Studies of directed thinking fall broadly into three main areas which are described as problem solving, reasoning, and decision-making. We shall consider each in turn.

**PROBLEM SOLVING**

A person has a problem whenever he or she wishes to achieve a goal and is unable to proceed immediately to do so. Problem solving consists of finding a method of getting from where you are to where you want to be, using such resources and knowledge as you have available. This definition obviously covers a vast range of human activity; problem solving is clearly involved in solving crossword puzzles and choosing chess moves, but it is equally involved in finding your way to a new destination, obtaining a ticket for a sold-out sporting contest, or working out how to persuade your boss to give you a pay rise.

One distinction which has helped psychologists think about the vast range of behaviours involved in problem solving is that between well-defined and ill-defined problems. In a well-defined problem, all the information needed and the means of solution are available at the outset. This is typical of things that are set as “problems” in newspapers, and so on, and also typical of much research in the psychological laboratory. An anagram is an example of a well-defined problem. You know the letters that constitute the solution word and also the means of solving the problem – rearrangement of the order of letters – at the outset. Well-defined problem solving thus consists of applying known rules to known information in order to transform the situation and achieve the goal.
Some of the most famous studies of well-defined problem solving were conducted by Newell and Simon (1972). An example of one of their problems is cryptarithmetic, in which subjects were given the following problem:

\[
\begin{align*}
\text{DONALD} & \\
+ \text{GERALD} & \\
\hline 
\text{ROBERT} & 
\end{align*}
\]

Subjects are also told that \( D = 5 \) and that each letter represents a single digit number between 0 and 9. Given this information and the assumption that the normal rules of arithmetic apply, it is possible – though complicated – to work out what all the letter-number pairings must be. If the reader wishes to attempt this problem, then it is suggested that a good record (on paper) of the sequence of attempts – including errors and correction – be kept.

Newell and Simon (1972) made an important theoretical contribution with the idea of problem solving as a search through a *problem space*. A problem space consists of a number of linked *states* including an initial or starting state and one or more goal states. All problems include *permissible operators* which allows one state to be transformed to another. Thus, solving problem consists in applying operators repeatedly to transform the initial state into a goal state.

As an example consider the game of chess (also studied by Newell & Simon, 1972). The states of the game can be described as the position of the pieces on the board plus some additional information (whose turn it is to move, do players have the right to castle, may a pawn be captured *en passant*, and so on). The initial state is thus the board with the pieces in starting position with White having the right to move. A goal state is any position in which the player has won the game either by checkmating the opponent or making such a mate inevitable. The permissible operators are the laws of chess, which determine the moves that can legally be made in a given situation.

Note that these definitions tell us nothing about the strategy of chess. The problem space consists of all states that can be reached by legal moves – a vast number of possibilities in the case of chess. The strategy of the game obviously consists in choosing between alternative legal moves in such a way as to move towards the goal state of a winning position. In chess, as in many other problems, the problem space is too large for an exhaustive search to be feasible. You cannot consider all moves and all possible replies to more than a very few moves ahead without the number of possible positions becoming enormous. Thus Newell and Simon (1972) emphasize the importance of *heuristic* strategies. An heuristic is a short-cut, rule of thumb method which may lead to a quick solution, but which may also fail. What heuristics do is to drastically reduce the size of the problem space to be searched in the hope that the goal state is not excluded in the process.
Consider the following anagram: GBANRIEK. Since it has eight letters the total problem space includes the $8! = 40,320$ possible rearrangements of the letters. A guaranteed, algorithmic (i.e., exhaustive search) method of solving this involves constructing all 40,320 letter strings and checking whether each is a word. A typical heuristic method, on the other hand, might involve looking for familiar letter patterns to decompose the problem. For example, we note that the anagram includes the letters I, N, and G and speculate that the word might be of the form ___ING. Thus we have now reduced the problem to solving the five-letter anagram BAREK which has only $5!$ (120) possible solutions and is thus much easier. We may now spot the solution word BREAKING. Like all heuristics, however, this was not guaranteed to work. Many words contain the letters ING in other configurations, e.g., GELATIN.

Problem space analysis is extremely useful as it provides a common framework in which to describe a very wide range of different problems. Newell and Simon (1972) studied subjects using think-aloud protocols while solving problems such as the cryptarithmetic example given above. They concluded that people have sets of general-purpose problem solving strategies that are used in similar ways to search problem spaces, no matter what particular domain is involved. They implemented their theory in a working computer program called General Problem Solver that was claimed to solve the same problems as the human subjects and in a similar way.

Important though this work has been, the conclusions are somewhat questionable. The first difficulty is that most real-life problems are ill defined. Some aspect of the problem – the information assumed, the means of solution, sometimes even the goal – is incomplete or missing at the outset. Take the case of engineering design which was subjected to detailed psychological study by Ball, Evans and Dennis (in press). An engineer is given a general specification for a device which includes its functionality – what it must do – and a number of constraints, including costs. The engineer must then come up with a technical specification for a device which can be constructed and can be demonstrated to work.

As Ball discovered, such problems are not at all well defined. Nearly all the information required to solve the problem is implicit and must be retrieved either from the existing knowledge and experience of the engineer or by researching technical manuals, and so on. In the process of design, constraints emerge that were not apparent at the outset. The goal initially set may also be modified and rethought as the work progresses. Now such activity can still be usefully described within the problem space framework – a space that is being continually augmented and redefined by the knowledge and experience of the engineer. However, the point is that simply applying the problem space description provides no explanation for some of the most important aspects of the process, particularly the means by which prior knowledge and experience are retrieved and applied.
A number of more recent studies of human problem solving have focused on ill-defined problems and the use of prior knowledge. Of particular interest has been the role of analogy in solving problems (see Gick & Holyoak, 1980, 1983; Keane, 1988). Most real-life problem solving — including “expert” problem solving — occurs within contexts where the solver has previous experience. Clearly, people do not solve all such problems as if seen for the first time; they must extrapolate from past experience. The theoretical and practical interest lies in how they actually bring their prior knowledge to bear.

A problem that has featured in many of these studies is the tumour problem first introduced by the Gestalt psychologist Duncker (1945). The problem is that of a patient who has a malignant but inoperable tumour that can be destroyed only by radiation. However, the radiation destroys healthy tissue at the same rate as diseased tissue. The solution that subjects must find is to use a lens to converge the rays at the point of the tumour. Hence, the rays accumulate only to sufficient intensity to destroy the tumour and not the healthy tissue they pass through on the way (see Figure 1).

The problem is incompletely defined in that while the goal and constraints are generally indicated, subjects must search their knowledge and imagination for possible means of solution. General knowledge of medical procedures is unhelpful; surgery is out by definition; drug treatments are of no relevance. The problem can, however, be facilitated by provision of a structural analogue such as the General story. The General is trying to attack a fortress which is well defended and which may be reached by a number of different roads. Each road is mined and may be safely crossed only by a small band of men. The General splits his force into small groups which approach

![Figure 1 Solution to Duncker's tumour problem](image-url)
simultaneously from different directions, and converge at the fortress with sufficient force to win the battle.

Gick and Holyoak (1980) showed that presentation of the General story could facilitate convergence solutions to the tumour problem provided that subjects were given a cue as to its relevance. There is a theoretical argument as to whether analogies can work by direct mapping of the elements of the analogy on to the problem, or whether the solution is mediated by an abstract schema. Gick and Holyoak suggest that subjects may construct and apply a convergence schema which is defined in terms of variables. For example, in the schema the goal is to destroy an obstacle, the means is a sufficient force, the constraint is that direct application is blocked, and so on. The General story could lead to development of a schema which is applied to the tumour problem.

The notion of schema is a useful one, in that it helps us to understand how knowledge may be abstracted, generalized, and applied in new situations. The notion will recur in the discussion of reasoning to which we now turn.

REASONING

Reasoning is the process of drawing conclusions or inferences from given information. An important distinction is that between deductive and inductive inference. Deductive reasoning involves drawing conclusions that are logically valid, that is, they necessarily follow from the premises on which they are based. Thus such inferences do not increase the amount of information contained in the premises; they merely render explicit what was previously latent information. The following are examples of valid deductive inferences:

- The television will work only if it is plugged into the mains;
- The television is not plugged into the mains,
- Therefore, the television will not work.

- John is taller than Jim;
- Paul is shorter than Jim,
- Therefore, John is taller than Paul.

The validity of the first example does not depend in any way on our knowledge of television sets, but only on our understanding of the connective “only if”. Any argument of the form \( p \) only if \( q \); not-\( q \), therefore not-\( p \) would be logically valid no matter what propositions we substitute for \( p \) and \( q \). Hence, validity depends on the form of the argument, not its actual content. In logic, the statement \( p \) only if \( q \) cannot be true in a world where \( p \) is the case and \( q \) is not the case. Hence, once we know that \( q \) is false we can infer that \( p \) must be false as well.

The second example requires us to know that the relation taller—shorter is transitive. A transitive relation is one where the objects are ordered in a single
line so that whenever $A$ is higher than $B$ on the scale, and $B$ is above $C$ then $A$ is also above $C$. Examples of other transitive relations are better–worse, warmer–colder, and darker–lighter. Many relations, of course, are not transitive. If $A$ is next to $B$ and $B$ is next to $C$ it does not follow that $A$ is next to $C$.

Deductive inferences are very important in intelligent thinking as they allow knowledge to be stored in generalities and then applied to particular situations. Thus if we want to watch television and discover one that is unplugged, we immediately plug it in. This is a simple example of reasoning in order to solve a problem. The limitation of deductive reasoning, however, is that it adds no new knowledge; thus we cannot learn by deduction. Induction is involved whenever our conclusion has more information than the premises. A typical example is an inductive generalization such as

The Australian soap operas I have seen were boring, hence all Australian soap operas are boring.

Such an inference is clearly not logically valid, though it could well influence what you watch when you get the TV plugged in.

The British psychologist, Peter Wason, invented two famous problems that have been used extensively to study both inductive and deductive reasoning. The inductive problem was first published by Wason (1960) and is known as the “2 4 6” task. The subjects are told that the experimenter has a rule in mind which applies to “triples” of three whole numbers. An example which conforms to the rule is “2 4 6”. The subjects are then asked to discover the rule by generating triples of their own. In each case the experimenter says whether the triple conforms or not. Subjects are told to announce the rule only when they are very sure that they know it.

The actual rule is “any ascending sequence” but the subject is induced by the example to form a more specific hypothesis, such as “ascending with equal intervals”. Most subjects have great difficulty in solving the problem initially because all the examples they test appear to conform to the rule. The reason is that subjects test positive examples of their hypothesis which invariably turn out to be positive examples of the experimenter’s rule as well. Their hypothesis can be refuted only by testing a negative example of the hypothesis such as “1 2 4” which is revealed as a positive instance of the actual rule. The set relationships involved are shown in Figure 2.

The protocols discussed by Wason (1960) were very interesting, suggesting that some subjects became so convinced of the correctness of their hypotheses that they were led to reformulate the proposed rule in different terms when told it was wrong. A striking example of this is shown in Table 1.

Wason’s interpretation of his findings was that subjects have a confirmation bias, meaning that they systematically seek out evidence that confirms rather than refutes their current hypothesis. He suggested that such a confirmation bias is a very general tendency in human thought which may
**Figure 2** Set relationships in Wason’s 2 4 6 task

<table>
<thead>
<tr>
<th>U</th>
<th>Universal set of all triples</th>
</tr>
</thead>
<tbody>
<tr>
<td>E</td>
<td>Experimenter’s rule – all triples in ascending sequence</td>
</tr>
<tr>
<td>S</td>
<td>Subjects’ hypothesis, e.g., ascending with equal intervals</td>
</tr>
</tbody>
</table>

*Table 1* Example protocol from Wason (1960)

- 8 10 12: two added each time; 14 16 18: even numbers in order or magnitude; 20 22 24: same reason; 1 3 5: two added to preceding number.

*The rule is that by starting with any number two is added each time to form the next number.*

- 2 6 10: middle number is arithmetic mean of other two; 1 50 99: same reason

*The rule is that the middle number is the arithmetic mean of the outer two.*

- 3 10 17: same number, seven, added each time; 0 3 6: three added each time.

*The rule is that the difference between two numbers next to each other is the same.*

- 12 8 4: the same number subtracted each time to form the next number.

*The rule is adding a number, always the same one, to form the next number.*

- 1 4 9: any three numbers in order of magnitude.

*The rule is any three numbers in order of magnitude.*

(17 minutes)
account for the maintenance of prejudice and false belief. While a number of authors have accepted this interpretation, it has also been subject to serious challenge (see Evans, 1989; Klayman & Ha, 1987).

The problem is that the subjects in the “2 4 6” experiment have no way of knowing that a positive test cannot lead to refutation of their hypothesis, and in many real-world situations it would do so. For example, in science it is customary to formulate general hypotheses and test if they apply to specific cases. Hence, given the hypothesis “All metals expand when heated” you would test any untried metal to see if the prediction holds – and if it did not you would indeed refute the hypothesis. You would not be likely to try heating non-metal things, and even if you did and they expanded, it would mean only that your rule was insufficiently general.

Arguments such as these have led some authors to suggest that subjects’ behaviour on the “2 4 6” is more rational than it at first appears and that if there is a bias, it is towards positive testing rather than to confirmation as such. A particularly interesting experiment reported by Tweney, Doherty, and Mynatt (1980) provides evidence for this. In one study, instead of defining instances in positive and negative terms (right/wrong, belonging/not-belonging) they told subjects that all triples were either MEDs or DAXes and that “2 4 6” was an example of a MED. What happened was that subjects continued to test their hypotheses positively but alternated between testing MED and DAX hypotheses. For example, if the hypothesis was that “triples ascending in equal intervals are MEDs and others are DAXes”, then they might test “1 2 5” predicting it to be a DAX. This meant that they effectively tested negative examples of the usual hypothesis and hence solved the problem much more easily. The psychological difference is that the negative test of MED was construed as a positive test of DAX.

A close parallel to these findings occurs with the second and most famous of Wason’s problems – the four-card selection task (see Evans, Newstead and Byrne 1993 for detailed review and discussion). This problem requires subjects to test hypotheses via deductive reasoning. In the classic “abstract” version of the task, subjects are told that a set of cards always has a capital letter on one side and a single-figure number on the the other side. They are then shown four such cards lying on a table with the exposed values as shown in Figure 3. The subjects are told that the following rule may be true or false:

*If there is an A on one side of the card then there is a 3 on the other side of the card.*

The subjects’ task is to turn over those cards – and only those cards – that are needed to decide whether the rule is true or false. The task is deceptively simple, since most subjects fail to solve it. The common answers given are A alone, or A and 3. The correct answer is the A and the 7. The reason is that the rule can be shown to be false only if there is an A on one side of a card and number other than a 3 on the other. Only by turning the A and the 7 (not a 3) is it possible to discover such a card. There is also no point
in turning the 3 since the rule makes no claim that an A must be on the back of a 3.

Wason's original claim was again that card selections reflected a confirmation bias: subjects were trying to prove the rule true rather than false, that is, looking for the combination A and 3, rather than A and not-3. This view was, however, refuted to the satisfaction of Wason as well as other authors by the finding of 'matching bias' reported by Evans and Lynch (1973). They pointed out that the preferred selections, A and 3, were not only the verifying choices, but also the positive choices matching the items named in the actual rule. Verification and matching could, however, be separated by introducing negative components into the rule. Consider for example, the rule

*If there is an A on one side of the card then there is NOT a 3 on the other side of the card*

If subjects have a confirmation bias, then they should now choose the A and the 7 which confirm the two parts of the rule. If, however, they have a matching bias then they should continue to choose A and 3 which are the correct and *falsifying* combination on this rule. Subjects do, in fact, continue to choose predominantly matching values on this and other variants of the rule, thus confirming the predictions of Evans and Lynch. Evans (1989) regards matching as an example of a generalized positivity bias, that is, bias to think about positively defined items, which also accounts for subjects' behaviour on the "2 4 6" task.

Dozens of experiments have been published — and continue to be published — in which subjects are asked to solve versions of the Wason selection task. Most of these have been concerned with the so-called thematic materials facilitation effect. This has its origin in two early studies discussed in Wason and Johnson-Laird's (1972) famous textbook on reasoning. In one of these (Johnson-Laird, Legrenzi, & Legrenzi, 1972) subjects were shown envelopes in place of cards, together with the following Postal Rule:

*If the letter is sealed then it has a 50 lire stamp on it.*

Subjects were then shown four envelopes which were either front side up and showing a 50 or 40 lire stamp, or rear side up showing that they were sealed or unsealed (see Figure 4). The subjects had to decide which envelopes to turn
over in order to decide if the rule was true or false. The usual matching response on the abstract task would lead to choice of the sealed envelope and the 50 lire stamp. However, almost all subjects made the logically correct choice of the sealed envelope and the one showing a 40 lire stamp.

The original interpretation offered of this and other similar experiments was that use of thematic materials facilitated logical reasoning on the task. This view has been considerably refined by subsequent research, however. The problem with the Postal Rule is that a very similar rule (involving pence rather than lire) was in force in the UK at the time of the study. Thus it was argued that subjects knew from experience that envelopes with a lower value stamp must not be sealed and that hence no “reasoning” as such was required to solve the problem. This argument was supported by the findings of several later studies which showed that first, the Postal Rule produces no facilitation of performance in American subjects unfamiliar with such a rule, and second, British subjects too young to remember the rule (it was dropped in the 1970s) show no facilitation on the problem whereas older subjects perform much better.

It is not the case, however, that subjects must have direct experience of the context in order for a problem content to facilitate on the selection task. A very effective version, for example, is the Sears Problem in which subjects
THINKING AND REASONING

are asked to play the role of a store manager checking that a company rule has been followed. The rule is

If a purchase exceeds $30, then the receipt must be approved by the departmental manager.

Subjects are shown four receipts, two of which are front side up showing totals of above and below $30 and two of which are front side down and either have or do not have the signature of the departmental manager on them. Few subjects have any difficulty in correctly deciding to turn over the receipt for more than $30, and the one that has not been signed by the manager. This is despite the fact that subjects have not worked as managers in department stores.

While arguments exist about the precise reason for facilitation of performance by these kinds of thematic content, the general idea is that where subjects have either direct or analogous experience that can be linked to the problem, then they can solve it. Another line of argument is that it is the introduction of deontic terms such as may and must which carry with them notions of permission and obligation that causes the facilitation. The idea is that we have generalized reasoning schemas that enable us to understand the logic of any situation in which, for example, a precondition is set for an action. Thus, once we have identified the action (e.g., sealing an envelope, spending over $30) and the precondition (sufficient value stamp, permission of departmental manager) we know what to do: we are applying a generalized permission schema to the problem at hand.

The two problems of Peter Wason discussed in this section have stimulated much interesting psychological work on the nature of human reasoning. The specific findings discussed here invite two general conclusions: first, that reasoning with “abstract” problem material is heavily biased by a tendency to think about positively rather than negatively defined information, and second, that the introduction of thematic problem content, and hence associated prior knowledge, can have a dramatic effect on the reasoning observed, and sometimes produces much better logical performance. The “sometimes” in the latter conclusion is needed. Other research, which there is no space to discuss here, has also indicated that prior knowledge can be a source of bias and error in reasoning. This is especially the case when subjects are asked to evaluate the logic of an argument but have strong prior beliefs about the truth of a conclusion (see Evans, 1989, chap. 4).

DECISION MAKING AND STATISTICAL JUDGEMENT

In a problem solving task, it is normally possible to work out and demonstrate a solution to the problem set. Once you have the solution, you know it and can prove it. In a decision-making task, however, subjects are required to exercise judgement about a choice that will only later prove to work out
well or badly. Decision-making means committing yourself to choices between actions by anticipation of what the outcomes will – or may – be. Thus when we make any decision – to accept one job rather than another, to marry someone or not, go to a football match rather than stay at home – we do so in the hope that the future we chose was to be preferred to the one we avoided.

Decision-making is obviously of great importance in the real world, but it is a subject of considerable psychological interest too. Most real-world decision-making is done under conditions of uncertainty: we do not know for sure what will happen as a result of each choice and at best can try to estimate the probabilities of different outcomes. If we are to choose rationally then we need to evaluate the desirability of these outcomes as well. In the parlance of decision theory, we should try to maximize expected utility where utility is the subjective value of the outcome and where the term “expectation” means that we weight the various possible outcomes by their likelihood of occurring. Hence, a small chance of a highly desirable outcome might be equally attractive to a much better prospect of a less desirable outcome.

There has been much debate in the psychological literature about whether people choose rationally or not. The notion of rational choice has several components. First, it implies that people will consciously consider the various actions available to them and try to project ahead the possible outcomes and further choices to which they lead in what is termed a decision tree. Second, it is assumed that they assign probabilities and utilities to each of these outcomes as accurately as possible in the light of their current beliefs. Finally, rational decision-makers are assumed to apply systematic principles, such as the maximization of expected utility, in order to decide their final choices.

There are many demonstrations of human choice behaviour that appear to depart from this idealized notion. Within the space restriction here I shall discuss just one aspect – the ability of people to judge probabilities or to reason statistically. A famous set of papers by the psychologists Amos Tversky and Daniel Kahneman dating from the early 1970s have apparently demonstrated the frailty of human probability judgement. This research is often cited as evidence of irrationality, although Tversky and Kahneman themselves follow the tradition of work on “bounded rationality” espoused by Newell and Simon (1972). The idea is that people cannot base their probability judgements on probability theory due its computational complexity and instead employ short-cut rules of thumb known as heuristics. While often useful, such heuristics can also lead to systematic errors and biases.

Of the heuristics discussed by Kahneman and Tversky, the two most famous are those of representativeness and availability (see Kahneman, Slovic, & Tversky, 1982 for a collection of relevant papers, including the seminal ones). Probability or frequency of an event is estimated by the availability heuristic when people base their judgement on the ease with which examples can be brought to mind. Such a heuristic would often be effective.
For example, an experienced doctor might base a provisional diagnosis on her recollection of the numbers of previous cases or patients with similar symptoms who turned out to suffer from a particular condition. Assuming that memory was accurate and that experience was representative then this is a good, if rough basis for a judgement.

As Tversky and Kahneman have demonstrated, however, relying on availability of recalled examples can lead to biases. For example, some types of information are easier to retrieve than others, due to the way in which memory is organized. For example, most people will say, if asked, that there are more words in English that start with the letter $k$ than those that have $k$ as the third letter, although the reverse is true. The problem is that it is hard to generate examples of the latter category: they cannot easily be "brought to mind".

Availability is also implicated in biases which preserve false beliefs and theories. An interesting example is the phenomenon of illusory correlation. It has been demonstrated in a number of studies that human judges — including experts — hold theories that are not supported by the evidence they encounter. For example, some clinicians maintain that projective personality tests such as the Rorschach ink blot test is useful in diagnosing mental illness despite a lack of any supporting evidence. Research has shown that such judges perceive a correlation between test results and diagnoses in a set of data in which they are in fact randomly related. A plausible explanation of illusory correlation is that the judges selectively remember the cases that confirm their expectations or pet theories. Thus confirming cases are more available in later recall and bias the judgement of the correlation.

The representativeness heuristic is involved in judgements of conditional probability. The likelihood of a sample given a population, or of an event given a hypothesis is dependent upon the perceived similarity of the two. Similarity judgements may, however, cause the subject to overlook the relevance of a critical statistical feature such as the size of the sample, or the base rate occurrence of the event. A simple example is provided by the conjunction fallacy (Tversky & Kahneman, 1983). Subjects are given a description of Bill as follows:

Bill is . . .

They are then asked to rank the likelihood of several statements including the following:

a Bill is an accountant
b Bill plays jazz for a hobby
c Bill is an accountant who plays jazz for a hobby.

What happens is that most subjects rate the order of likelihood of these statements as $a > c > b$. However, there is a statistical impossibility here in that statement $c$ cannot be more likely than statement $b$. Given two events $A$ and
B the probability of them both occurring – \( P(A \cap B) \) – must be less than or equal to the probability of either \( P(A) \) or \( P(B) \). Whenever \( c \) is true then \( b \) is true as well, because Bill plays jazz for a hobby. If all jazz players were accountants then the two statements would be equally likely, otherwise \( b \) has to be more probable.

The explanation offered for the fallacy is that the description of Bill conforms to our stereotype for accountants but not for jazz players. Thus the statement \( c \) is more representative of the description than is statement \( b \) and hence judged more probable.

One of the most famous of Kahneman and Tversky’s problems is the Cabs Problem. You are given the following information: in a certain city there are two cab companies: the Blue cab company, which has 85 per cent of the city’s cabs, and the Green cab company, which has 15 per cent of the city’s cabs. A cab is involved in a hit-and-run accident and a witness later identified the cab as a Green one. Under tests the witness was shown to be able to identify the colour of a cab correctly about 80 per cent of the time under comparable viewing conditions. The subjects are asked if the cab involved in the accident is more likely to have been Green or Blue. Most say Green, although the correct answer is Blue.

The problem is that subjects disregard the base rate or prior probability of the cab colour – 85:15 in favour of Blue. In fact, when asked to give a numerical estimate, most subjects say 80 per cent Green – the chance of the witness correctly identifying a cab. If there were no witnesses, it would be obvious that the chance of the cab being Blue was 85 per cent – the base rate. As Figure 5 shows, however, the chance of a Blue cab being identified as Green is 17 per cent which is still higher than the chance (12 per cent) of a Green cab being identified as Green.

\[
\begin{align*}
A & : \text{Probability of Blue identified as Blue} = 80\% \times 85\% = 68\% \\
B & : \text{Probability of Blue identified as Green} = 20\% \times 85\% = 17\% \\
C & : \text{Probability of Green identified as Green} = 80\% \times 15\% = 12\% \\
D & : \text{Probability of Green identified as Blue} = 20\% \times 15\% = 3\%
\end{align*}
\]

Figure 5 Probabilities in the Cabs Problem
Originally, the base rate fallacy was interpreted as the base rate lacking representativeness, although the explanation is probably more fundamental. We find it very difficult to apply abstract statistics to individual cases. Hence, many cigarette smokers are aware of the statistical risks for smokers as a whole, but do not feel that this affects them as individuals. However, we can apply statistics when we see a causal connection. If the cabs problem is slightly reworded, most subjects give the right answer. In this version the number of Green and Blue cabs in the city is the same, but 85 per cent of the cabs involved in accidents are Blue. The image of reckless Blue cab drivers conjured up induces subjects to take account of the base rate, although from a statistical point of view the problem is unchanged.

CONCLUSIONS

Psychological research on thinking and reasoning has produced some useful — and sometimes surprising — conclusions. The common-sense view, that intelligent actions are based on conscious and rational acts of thinking, does not fit the evidence at all well. If thought is to be defined as the information processing that underlies problem solving, reasoning, and decision-making, then surprisingly little of this appears to be accessible through introspection.

If human thinking is rational — and the success of the species suggests that it should be — then that rationality is highly constrained by our capacity to process information. In particular, we seem to solve problems and make decisions largely on the basis of heuristic processes which serve us well in some circumstances, but lead us into error and bias in others. We seem to have particular difficulty in understanding probability and uncertainty despite the crucial role that this plays in rational decision-making.

Studies of reasoning also show that we are prone to biases, for example in a strong preference for thinking about positively defined information. Perhaps the most important finding in this area, however, is the discovery that we do not — as was once thought — appear to reason by the use of an abstract mental logic, but instead seem to be highly influenced by the content and context of the problems with which we are faced. The processes of human thought appear to be quite specific to the areas of knowledge which we are involved in applying.

FURTHER READING


**REFERENCES**


Artificial intelligence, almost always known as AI, attempts to understand intelligent behaviour, in the broadest sense of that term, by getting computers to reproduce it. “Intelligent behaviour” is taken to include thinking, reasoning, and learning, and their prerequisites (perception, the mental representation of information, and the ability to use language). Indeed, much current work in AI is concerned with modelling aspects of behaviour that would not normally be thought of as requiring any special intelligence. As part of computer science, AI is separate from cognitive psychology, although there is a large overlap in subject area. The two come together (with, most importantly, linguistics and philosophy) in the multidisciplinary approach of cognitive science.

Although AI aims to understand human intelligence, it also aims to produce machines that behave intelligently, no matter what their underlying mechanism. However, although these machines may not model human behaviour, their construction may reflect principles that are useful in studying it.
HISTORY

Since AI depends on computers, it is a relatively new discipline: the name was first used in the mid-1950s, though a few years earlier, pioneers such as Alan Turing in Britain and Claude Shannon in the United States had worked out how to write chess-playing computer programs. The dream of mechanized thought has, of course, a much longer history. The philosophers Blaise Pascal (1623–1662) and Gottfried Leibniz (1646–1716) built small calculating machines, and conceived grander schemes for formalizing thought processes. Charles Babbage (1792–1871) came nearer to building a universal computing machine, but was foiled by the limitations of having to use mechanical parts. Real computers had to wait for electronic components — first vacuum tubes, then semiconductors.

A conference at Dartford College, New Hampshire, in 1956 effectively launched AI research, even though its organizers felt disappointed at the time. In retrospect, the most important line of research discussed at the conference was that of Allen Newell and Herbert Simon (see e.g., Newell, Shaw, & Simon, 1957) on human problem solving. They proposed the idea of a heuristic ("rule-of-thumb") procedure for solving problems, and they shunned a line of research based on modelling the properties of networks of brain cells, which only assumed major importance again 25 years later, in the guise of connectionism. Newell and Simon's information processing approach was the dominant one in the early days of AI, and it remained influential throughout the 1960s — the so-called semantic information processing era. There was, however, a subtle shift of emphasis from a formal analysis of tasks to one based on the meaning of the information being processed. Furthermore, in attempting to tackle broader problems, such as natural language understanding, AI researchers quickly discovered that everyday tasks depend on huge amounts of background knowledge. To keep programs manageable, they were made to work in limited domains, in particular BLOCKSWORLD — a tabletop with prismatic blocks on it. It was hoped that programs that worked in these limited domains would scale up to real situations. In practice they did not, and in retrospect it is often obvious why they could not.

The 1970s was a somewhat disappointing period in "traditional" areas of AI research. Indeed, in the UK the Lighthill report (Lighthill, 1972) concluded that AI should not be a priority area for research. The late 1970s saw four important developments. The first was a shift in interest from specific computer programs to general principles. To some extent this development was linked to the second, the emergence of cognitive science, in which AI techniques are used with the primary goal of developing general theories of cognition, rather than with the more applied ("engineering") goal of building intelligent machines. The third development was a shift in the research topics seen as central to AI. In particular, fifteen years of research on the first expert systems was beginning to have spectacular payoffs (in the domains of
mathematics, medical diagnosis, and determining the structure of complex organic molecules) and suddenly everyone wanted to write an expert system. In the short term, this enthusiasm generated additional funding and research, but it soon became apparent that an expert system in one domain could not necessarily be used as a model for one in another domain. If expert systems showed that real applications had to come to grips with formalizing real knowledge (as opposed to knowledge about toy domains), they also showed that this task was a formidable one. The fourth development was the re-emergence of neural network modelling, of the kind that had been largely set aside by those who espoused the Newell and Simon information processing approach. Theoretical developments together with the availability of larger, faster computers suddenly saw this approach producing important and enticing results.

The 1980s saw the working out of these developments. Although all remain important, all have faced disappointments. It is very hard to make an expert system that replaces an expert, though much easier to write a program that helps one. And it is hard to generalize the lessons learned in one domain of expertise. Cognitive science has not integrated its subdisciplines as closely as was hoped, and neural network modelling has still to show that it can make significant contributions to modelling abilities that call for complex information processing, in particular high-level processes in language understanding and thinking and reasoning.

**KNOWLEDGE REPRESENTATION**

Intelligent behaviour requires information to be stored, either in a short-term store or a long-term store or, more usually, both. One of the primary tasks of AI is therefore to produce an account of how information is represented in an intelligent system.

We know that the human nervous system has many parts, and that those parts probably operate in different ways. Nevertheless, there are many attractions in proposing that all information is stored in the same format. It may not be the form of information storage that differentiates information processing systems, but the nature of the information and the purpose for which it is used. Partly for this reason, many AI researchers have been attracted to the idea that information should be stored using the logical language known as *first order predicate calculus* (FOPC), and extensions of it that incorporate reasoning about time and modality. An additional attraction of this proposal is that, at least in principle, FOPC is computationally tractable: given a FOPC database, other facts implied by that database can be generated automatically. Other systems of representation are either not known to have or known not to have this property.

Unfortunately, although FOPC appears to have desirable properties, in practice it is extremely cumbersome to use. Partly because of the uniformity
of the representation, facts in a large FOPC database can be difficult to find. Similarly, although there is a well-established procedure for drawing inferences from facts in a FOPC database (the resolution method, Robinson, 1965), it very quickly gets bogged down in making all but the simplest inferences. Furthermore, inferences made from a FOPC database cannot be overridden by new information. Everyday inferences can — they are said to be non-monotonic. For example, if I know that John is 25 years old and lives in Los Angeles, I infer that he can drive. If I subsequently learn that he suffers from epilepsy, I would probably withdraw my previous conclusion. Since the late 1970s there have been several attempts to construct non-monotonic logics, similar to FOPC but with additional rules of inference that violate monotonicity. There have also been attempts to formalize non-monotonic reasoning in other ways. The idea of a truth maintenance system (TMS) (Doyle, 1979) has been important in many of these. A TMS stores information about the justification for beliefs held, and allows dependency-dependent backtracking, so that when a belief turns out to be false, the reasons why it was held can be accessed directly and reassessed. None of these attempts to handle non-monotonic reasoning has been entirely successful.

Partly as a result of problems with uniform representation systems, such as FOPC, many AI researchers have proposed non-uniform representations, which allow special procedures for manipulating certain types of information. One of the earliest, and best-known, non-uniform representations is semantic networks (Quillian, 1968). Semantic networks give a special place to the information represented in their links and, in particular, they allow efficient processing of taxonomic information. Quillian’s original, and rather simple, networks have been extended and elaborated in various ways, and representation of information in network form has proved a recurrent theme in AI. More complex non-uniform representation schemes that are related to semantic networks include frames and scripts. Scripts represent stereotyped sequences of events, frames have several uses. In one, frames represent particular objects and types of object, and a more recent development is that of object-oriented programming languages. The first widely used object-oriented language was the AI language SMALLTALK. More recently object-oriented versions of the most important AI language, LISP, have appeared, and languages such as C now have object-oriented versions (C++). Indeed, one of the major applications of object-oriented programming is not in AI, but in the development of windows-based interfaces for personal computers and workstations, where windows are treated as objects.

In the framework of semantic networks, the spread of activation through a network is the principal method of extracting information from it. This process has usually been simulated on a serial computer, but it ought to be achieved more efficiently on parallel hardware. Indeed, one of the most important parallel processing computers, the Connection Machine (not to be confused with connectionist neural nets), was inspired by Scott Fahlman’s
(1979) suggestion for implementing semantic networks on special hardware. The idea of distributed processing is also found in neural network models of cognitive processing. Neural networks also allow, though they do not demand, distributed representations of the knowledge embodied in them. In particular, those neural networks that learn to perform tasks, rather than having information encoded into them by the programmer, are likely to develop distributed representations. Such networks show rule-governed behaviour as an emergent property, and the only way to determine exactly what rules such a network is following is to examine the relation between its inputs and its outputs.

There are many things we cannot be sure of, so a further issue in knowledge representation is the encoding and use of uncertain information. Inferences from uncertain information are modelled mathematically using probability theory and, in particular, Bayes’ theorem, which is familiar to psychologists from statistical courses. Complex sets of probabilistic interrelations can be modelled in so-called Bayesian networks. Unfortunately Bayesian inference is neither computationally simple nor always the correct model of real world uncertain inference. The early expert system MYCIN (see below) introduced the simplifying idea of certainty factors associated with each of its diagnostic rules of inference. In recent years attention has focused on a more sophisticated mathematical approach known as Dempster-Shafer theory and there has also been renewed interest in fuzzy set theory, which enjoyed brief popularity in cognitive psychology in the mid-1970s.

VISION

Traditional AI research on vision was concerned, broadly speaking, with recognition of the objects – the prismatic solids – in the BLOCKSWORLD. For computer vision programs, the objects were matt white, uniformly lit (no shadows), and placed against a black background. In fact, the general problem of object recognition in the BLOCKSWORLD was set aside in favour of two of its component problems: finding lines in an image of a BLOCKSWORLD scene, and segmenting the image into sets of regions – each region corresponding to a surface – that belong to the same object. Indeed, this research came to be dominated by attempts to solve the segmentation problem: many programs required line drawings (rather than images) as their inputs.

The most important method of attempting to solve the segmentation problem, originally suggested by Alfonso Guzman (1968), was to use information about the types of vertex in the scene. Guzman’s taxonomy was intuitive, but it was systematized independently by Max Clowes (1971) and David Huffman (1971), who stressed the importance of maintaining different descriptions of the image (in terms of lines, line junctions, and regions) and the scene (in terms of edges, vertices, and surfaces), and of making systematic
ARTIFICIAL INTELLIGENCE

inferences about the scene on the basis of the image. The Clowes-Huffman scheme is limited to scenes with no shadows and in which no more than three lines meet at any point. It has three types of line (corresponding to boundaries, inside edges, and outside edges) and four basic types of line junction (Ts, Ys, Ls, and arrows). From these line types and junction types, 16 derived junction types can be constructed, which correspond to possible

Figure 1  The 16 derived junction types in the Clowes-Huffman scheme – 4 Ts, 3 Ys, 6 Ls, and 3 arrows. An arrow on a line signifies that it represents an occluding edge (boundary between objects), a plus (+) sign signifies a convex (or outside) edge of a single object, and a minus (−) sign a concave (or inside) edge. The direction of the arrow indicates the side of the line on which the occluding object lies (to the right when facing in the direction of the arrow)
configurations in a BLOCKSWORLD scene (see Figure 1). Identification of the basic junction types in the image, plus the application of the constraint that any line should be of the same type along its whole length, allows most images of permissible scenes to be interpreted.

David Waltz (1975) extended the Clowes-Huffman scheme to scenes with shadows and to images in which more than three lines meet at a point. These apparently simple changes increased the number of permissible derived junction types from 16 to about 2,500. Nevertheless, Waltz’s program was more successful than those devised by Clowes and Huffman, since he exploited the need for consistent labelling of neighbouring junctions. An iterative technique known as Waltz filtering or, more generally, as relaxation eliminates possible labellings of junctions, using this consistency constraint. In most cases it rapidly converges on a solution to the segmentation problem for the image it is processing.

Steve Draper (1981) and others have identified a number of problems with the junction-labelling technique and with an alternative to it known as the gradient-space method. Draper invented a technique called sidedness reasoning. Sidedness reasoning is about whether two points or surfaces are on the same side of a third surface. Draper showed that this technique was able to segment all BLOCKSWORLD images but in doing so he virtually put a stop to work on object recognition in the BLOCKSWORLD. The reason was that his technique wore on its sleeve the fact that it was specific to BLOCKSWORLD: it works only when all surfaces are flat. Thus, the idea of solving the problem of object recognition in a miniature domain and scaling up the solution to the real world would not work.

A quite different approach to the problems of vision is found in the work of David Marr (1982) and his associates. Marr’s work integrates ideas from AI, psychology, and neurophysiology in what is usually taken to be the paradigmatically successful piece of research in cognitive science. The work is guided by an underlying philosophy about the study of natural information-processing systems. Marr identified three levels at which such systems should be studied. First, a task analysis answers the questions of what the system does and why it does it. This analysis leads to a computational theory of the system — an account of the function (in the mathematical sense) it computes. The second level of analysis is that of representation and algorithm. The third level is that of implementation. In the case of natural information processing systems, this level of analysis requires the study of the neural mechanisms that support the system. Marr is critical of previous AI work on vision, largely because of its focus on the second level of analysis at the expense of the first, to which Marr attached great importance. He is also critical of neurophysiological work, such as that of Hubel and Wiesel (1962), in which the purpose of certain types of cell is inferred from their properties. According to Marr, the purpose of a system (and of its parts) can be determined only by constructing a computational theory.
In his own work, Marr recognized three main stages of visual processing. In the first of these stages, the array of light falling on the retina is transformed into a representation called the *primal sketch*. The primal sketch is a symbolic representation, but it is a representation of the image, not of the scene. It contains information about lines, boundaries, and regions in the image. The construction of the primal sketch takes place very early in the visual system and proceeds on the basis of local interactions between processing units (cells) that represent adjacent parts of the image. Although these interactions reflect what is known about the early visual system, Marr eschewed theories that were motivated solely by neurophysiological evidence. Hence, his demand for independent support — from task analysis and psychological evidence — for the algorithm and representation he proposed.

In the second stage of visual processing, the $2\frac{1}{2}$D sketch is derived from the primal sketch. This sketch is a very short-term memory store into which a set of processes writes information about the surfaces (in the scene) represented in the image, their orientation, and their approximate distance from the viewer: the third dimension is not properly represented, hence $2\frac{1}{2}$D sketch. The most important of these processes are stereopsis, structure from motion, and shape from shading.

Since objects have not yet been recognized, surfaces cannot be identified by reference to information about the objects of which they are part. This aspect of the construction of the $2\frac{1}{2}$D sketch reflects Marr's preference for bottom-up (data-driven) theories of visual processing. The only world knowledge that such theories can claim the visual system uses is a set of general principles, such as what very few points in an image correspond to abrupt changes in the surface represented. Specific information about the scene being viewed is not yet available.

In the final stage of visual processing, a *3D model description* is constructed from the $2\frac{1}{2}$D sketch. This representation contains information about the identity and three-dimensional structure of the objects in the scene. Marr's account of this final stage is highly speculative, and less closely linked with the psychological and neurophysiological evidence. Marr's basic idea is that objects can be represented, in a catalogue stored in long-term memory, as jointed generalized cylinders (cylinders whose cross-section changes along their length). The principal axes of these cylinders make up stick figures of the objects represented. He showed that, subject to certain constraints, generalized cylinder representations could be derived from the $2\frac{1}{2}$D sketch, and then compared with entries in the catalogue, with any necessary rotation and bending at the joints. In practice this matching is difficult, and Marr suggested a process of gradual refinement in the match between the image and the stored representations in the catalogue. This kind of process can be (relatively) time-consuming, and was rejected by Marr in his analyses of the lower levels of visual processing.

365
Marr’s work incorporates, in addition to traditional AI-style programming, much straightforward mathematics. Subsequent work on vision, both theoretical and applied, has become increasingly mathematical and, hence, increasingly inaccessible to psychologists. On the theoretical side, many of the problems of visual analysis have been identified as special cases of what are known as ill-posed problems. They are ill posed because, as they stand, they do not have a unique solution. They can be analysed by a technique known as regularization, which requires the addition to the problem of the kind of general constraints identified by Marr. On the applied side, specialized hardware in the form of very large-scale integration (VLSI) chips has allowed, for example, stereo algorithms to be used in real-world applications.

THINKING, REASONING, PROBLEM SOLVING

Historically, problem solving was one of the earliest topics of AI research. Furthermore, it has often been argued that it is the central topic, since AI techniques in other domains can be seen as special cases of searching through a “space” of possibilities for a solution to a problem. For example, parsing a sentence can be seen as a search through the (infinite) set of possible syntactic structures defined by the grammar of a language.

Occasionally it is possible to examine all possible solutions to a problem to find the right one. However, for most interesting problems there are too many possibilities to make this approach viable. Usually there are several steps in the solution to a problem, so the number of possible moves multiplies up at each step, producing what is called a combinatorial explosion in the number of potential solutions. A control strategy for searching through the space of possible solutions is, therefore, required.

Traditionally, there are two ways of representing problems so that a search can be made for their solution. In a state-space representation, problems are represented in terms of states of the relevant part of the world, and actions (usually referred to as operators) that transform one state into another. In this representation, a single path through the tree of possibilities (= a sequence of operators) represents the solution to the problem. In a problem-reduction representation a large problem is broken up into a number of sub-problems, all of which must be solved if the main problem is to be solved. State-space representations are easier to construct. Sensible reductions of problems can be hard to find, but they are very useful when they have been found. In serious AI work on problem solving the two types of representation are combined into AND/OR trees. AND branchings represent problem reductions, where all the sub-goals have to be fulfilled. OR branchings represent alternative possibilities in a state space, only one of which has to be fulfilled.

Various general control strategies for searching problem spaces have been proposed. The most fundamental distinction is between breadth-first and
depth-first search of trees. In depth-first search one possible solution is followed up until it succeeds or fails, or until a pre-set depth limit is reached, since a branch in an AND/OR tree may never terminate. In depth-first search all possible one-operator solutions are checked, then all possible two-operator solutions, and so on. In depth-first search one possible solution is followed up until it succeeds or fails, or until a pre-set depth limit is reached, since a branch in an AND/OR tree may never terminate. Simple depth-first and breadth-first search are used only in desperation. Usually some method is introduced for following up the most promising possibilities. Methods for deciding which possibility is the most promising are inevitably heuristic. The most sophisticated method of making the choice is the AO* algorithm. However, the algorithm itself does not provide the means of measuring which next move is the best. Furthermore, there is no general method for assigning values to moves. A new one must be devised for each domain in which the algorithm is used.

Such methods can, nevertheless, be applied to solving puzzle-book problems and in game-playing computers (e.g., for chess). In chess-playing programs the problem that the computer is trying to solve is not how to win the game, but what move to make next. Successful programs run on very fast super-computers, so that they can examine vast numbers of possible moves. However, they limit the distance ahead (in terms of moves) that they look. Since they typically cannot see ahead to a winning position, they have to evaluate the positions that they can reach in other ways, and then aim to reach the best position that a rational opponent will let them. The play of such programs differs in several ways from that of human chess players. The standard of the best of them, however, is usually reckoned to be in the grandmaster category.

Even if all AI researchers had access to the kind of super-computers that chess programmers use, they would not necessarily want to use the same kind of brute force problem solving methods, particularly if they were interested in modelling human problem solving abilities. Newell, Shaw, and Simon (1957) first introduced the idea of heuristic (rule-of-thumb) problem solving techniques in their Logic Theory Machine, that proved theorems of logic. An alternative way of speeding up problem solving is to use domain-specific techniques, that may be heuristic, but which need not be. An early example of an AI program that used a domain-specific technique was Gelernter's (1963) Geometry Machine, which constructed the equivalent of geometrical diagrams. It is thought that most human mathematicians, except when they are working in completely new areas of mathematics, use domain-specific techniques. More generally, domain-specific techniques are thought to be widely used in all types of problem solving.

**LANGUAGE**

There is a long history of computational research on all aspects of language processing. Research on speech, both automatic speech recognition and speech synthesis, has been strongly influenced by work on signal processing
carried out by electronic engineers. More recently, with the advent of larger and more powerful computers, the field of *speech and language technology* has emerged, which is primarily directed to producing tools for processing large corpora of linguistic data held on computers. Some of the techniques developed may be of interest to AI researchers; others are used to derive statistical information that is of primary interest to, say, lexicographers.

Work on language processing is divided into three parts, concerned respectively with recognizing or selecting words, computing or generating sentence structure, and processing meaning at the level of discourse. Until the 1970s AI research on language processing often produced working systems that understood a substantial portion of a language such as English. Winograd’s (1972) SHRDLU, a program that talks about moving blocks around the BLOCKSWORLD, represents the apotheosis of this work. However, it has since become obvious that the component parts of language processing are each so complex that they must be studied separately, if real progress is to be made.

Recent work on word identification has been largely dominated by neural network modelling, in particular the TRACE model of auditory word identification (McClelland & Elman, 1986) and Seidenberg and McClelland’s (1989) model of visual word identification. The TRACE model is “hand-coded”. It does not use distributed representations, and hence its mode of operation is easy to discern. It has interacting banks of detectors at three levels: for the auditory features of sounds, for phonemes (sounds that correspond roughly to letters), and for words. The Seidenberg and McClelland model, on the other hand, is a model that learns. One of its most interesting features is its eschewal of lexical representations: all its knowledge is encoded in links between orthographic and phonological features.

Investigations of the computation of sentence structure (parsing) have taken two rather different directions. On the one hand, computational linguists worry about problems such as the linguistic niceties of describing sentence structure and the computational properties of the procedures that derive the structure for a particular sentence, given a description of how sentences in its language can be structured (a grammar). One of the most important developments in computational models of parsing is the introduction of unification-based approaches (e.g., Kay, 1985). Unification is a technique that is widely used in other branches of AI, in particular theorem proving. Unification-based parsers, like some other parsers, such as chart parsers, have the additional advantage of clearly separating information about how sentences can be structured (the grammar) from information about how sentence structure is computed (the parsing algorithm). In contrast with researchers whose primary interest is in the computational properties of parsing systems, those who attempt to model the way that people derive sentence structure have to take account of well-established empirical findings on, in particular, what happens when people encounter a syntactic ambiguity. It is not yet clear how these two approaches to parsing can be integrated.
Understanding and generating discourse still remain formidable tasks. AI research has often been hampered by a restricted or ad-hoc approach to word meanings. One hope is that linguistically more sophisticated approaches to word meaning, such as Jackendoff’s (1990) conceptual semantics, will be taken up by AI researchers. At the level of sentence meaning, AI researchers, at last, agree about the importance of compositional semantics of a broadly Montagovian kind (Dowty, Wall, & Peters, 1981). However, the major problems in describing discourse level processing, which have been known for many years, still resist satisfactory analysis. Some of the most important are figurative and indirect uses of language, coherence, ellipsis, and the role of the other participants’ beliefs.

LEARNING

For historical reasons, learning has been a comparatively neglected topic in AI. The information processing approach to understanding intelligent behaviour was seen as a radical alternative to the behaviourism that had dominated psychology, and which placed a strong emphasis on learning. Furthermore, traditional AI aimed to study intelligence at an abstract level, independent of both its genesis (learned or programmed) and its underlying mechanism (carbon or silicon). The study of learning has come back into its own with the increasing importance of connectionist modelling. Nevertheless, a number of important studies of learning have been carried out in the symbolic framework, and the diversity of the learning mechanisms that they investigate contrasts sharply with the behaviourist approach.

Learning by being told often involves little more than adding a fact to a database. However, more abstract pieces of information, such as advice on the best strategy for winning a game, may need to be operationalized.

A more complex kind of learning is learning from mistakes. Gerald Sussman’s (1975) program HACKER writes its own mini-programs for solving problems of stacking and unstacking blocks in BLOCKSWORLD. However, it can learn only when it can almost solve a problem, and its performance is crucially dependent on its having a “teacher” who presents it with a suitably graded set of problems. Patrick Winston’s (1975) program that learns concepts for configurations of blocks (such as arches) in BLOCKSWORLD, similarly learns from almost correct information. When told that something is not quite an arch, it can use that information to deduce what distinguishes arches from non-arches.

As well as recognizing the importance of being almost correct, Winston also emphasized that an important aspect of learning is what is sometimes called induction – going beyond the information embodied in the examples presented to the program to form general concepts (in his case) or rules. Positive instances suggest generalizations of the concept or rule, negative instances suggest specializations (or restrictions). Research subsequent to
Winston's, particularly that of Ryszard Michalski (e.g., 1983) has systematized the study of induction, and shown that it can be regarded as a special case of search, with the search space being the set of possible generalizations statable in a particular language. Michalski's approach is more powerful than Winston's, but less closely related to human learning. It can also be used for the related task of discrimination learning. Its disadvantage is that it works straightforwardly only if the generalizations are formulated using exactly the same predicates that are used to describe the instances.

Winston's program can learn more complex concepts (such as arch) only because it knows simpler concepts (pillar, lintel). This aspect of the program relates, very crudely, to the question of how much of what we know about language is learned, and how much is innate. In the case of concepts, it has been argued (e.g., by Fodor, 1981) that all concepts must be innate. More generally, it is widely, though not universally, believed that many general principles governing the form of possible languages are innate, and that the availability of these principles to the language learning mechanism explains how it is able to achieve what appears, on mathematical analysis, to be a difficult or impossible task.

Another famous example of learning by generalization is Arthur Samuel's (1963) checkers (draughts) program. This program develops a general method for evaluating board positions by comparing computed evaluations with the way the game actually turns out, and revising, if necessary, the method of evaluation.

A more ambitious, and more controversial, attempt to study a different kind of learning – learning by exploration – is found in Doug Lenat's (1982) AM (Automated Mathematician) and EURISKO programs. AM starts with a collection of set-theoretic concepts and ways of combining them, and creates further mathematical concepts from them (e.g., positive whole number, prime number, the fundamental theorem of arithmetic – that every number can be expressed as a product of prime factors).

None of the programs described so far provides a convincing model of human learning. People can learn things very quickly, though they often make mistakes in doing so. This very quick learning depends on particular ways of using background knowledge. Two lines of research that attempt to model this kind of learning investigate analogy-based learning and explanation-based learning. The importance of analogy in learning and problem-solving has long been recognized in cognitive psychology. None the less the underlying processes are difficult to model computationally, not least because the domain from which an analogy is drawn need not be specified in advance. In explanation-based learning (see e.g., de Jong, 1988) a single event or episode is explained on the basis of a theory about the relevant aspects of the world. That explanation is then generalized so that it will be useful in other situations.

Traditional AI work on learning has embodied a variety of ideas. An
alternative tradition, running from the British Empiricist philosophers of the seventeenth and eighteenth centuries to the behaviourists and neo-behaviourists of the twentieth century, has seen all learning as the formation and strengthening of associations between ideas. In a modified form, this notion also underlies recent connectionist accounts of learning. Connectionists machines are collections of simple processing units, with levels of

*Figure 2*  A simple connectionist network showing the three types of unit – input, hidden, and output – and the connections between them
activation that can be passed from one unit to another. A typical machine has three layers of units: input units, hidden units and output units (see Figure 2). Such machines can learn in several ways, but the most popular is known as a back propagation. It is a supervised learning method in which a stimulus is encoded at the input units and produces an output at the output units. The supervisor tells the machine what the output should have been, and the difference between the actual and expected outputs is propagated back through the network of units, and used, in a precisely specified way, to adjust the (associative) strengths of the connections between them. Adjustments are small, because the machine must not produce the correct response to the last input at the expense of responding grossly incorrectly to other inputs. Learning is slow, sometimes very slow, but a stable set of associative strengths is usually reached.

Another biological metaphor that has inspired AI work on learning is evolution. Genetic algorithms (e.g. Goldberg, 1989) use complex rules to perform tasks. The parts of these rules can be recombined by processes that are analogous to the genetic operations that take place in the germ cells during sexual reproduction. The resulting rules are then allowed to perform their task for some time, and their performance is assessed. Those that do best re-enter the "reproductive" process.

APPLICATIONS

Intelligent machines should be of more than academic interest. However, most of the machines that we interact with in everyday life, for example automatic bank tellers, are not intelligent. More intelligent machines — often referred to as expert systems — do have applications. However, despite the hopes of the early 1980s, it now appears that expert systems will typically be used to assist experts, rather than to replace them. Perhaps the most important area of application for intelligent programs is in medical diagnosis, though there are obviously ethical problems in this domain. One area in which computers play a crucial role is in modern scanning techniques (CAT, PET, NMR, etc.). The basic use of computers in scanning is to generate appropriate images. Intelligent programs might also help to produce diagnoses from images.

One of the earliest, and best known, medical diagnosis systems is MYCIN (Shortliffe, 1976), which diagnoses serious bacterial infections so that life-saving antibiotic drugs can be administered before a culture has been developed. The development of such a system requires the gleaning of information about the diseases in question and their symptoms. Some of this information is elicited from experts, sometimes with difficulty, as the experts cannot necessarily verbalize their knowledge. TEIRESIAS (Davis, 1982) is a program that attempts to automate this knowledge transfer, and also to use the knowledge already in MYCIN to generate user-friendly explanations of
ARTIFICIAL INTELLIGENCE

its diagnoses. Other diagnostic information comes from statistical records. In an expert system all the information is usually represented in a uniform way, so that new information can readily be added. The rules for making inferences are stored separately, and an attempt is made to keep the inferential processes simple. One of the major aspects of inference in expert systems is combining uncertain bits of information to produce a best guess, for example at a diagnosis. This combination is sometimes achieved using standard statistical (Bayesian) techniques and sometimes using domain specific rules, as in MYCIN (see above).

MYCIN also formed the basis of the first expert system shell, E-MYCIN, which is MYCIN stripped of its domain-specific knowledge. Expert system shells were the first of several attempts to make the creation of new expert systems easy. Success has been partial. E-MYCIN, for example, is most successful in other medical diagnosis systems, such as PUFF, which diagnoses pulmonary diseases.

Another well-known expert system is DENDRAL (Lindsay, Buchanan, Feigenbaum, & Lederberg, 1980), which works out the molecular structure of large organic molecules from their mass spectrograms. DENDRAL has been in regular use by research chemists for some time. An additional program, meta-DENDRAL, attempts to formulate new rules using the induction techniques described above.

A second area in which AI has sought to find application is in computer-assisted learning (CAL). With the expansion of higher education in the UK, CAL is likely to become increasingly important, though it is as yet unclear what the contribution of AI techniques will be. The current focus of attention is on multimedia, and in particular hypermedia learning tools, which provide facilities for exploring large databases in various ways, but which rely on much of the intelligence resting in the instructions and with the student.

The intelligent tutoring systems of AI, on the other hand, try to be intelligent themselves. Such systems have three main components: a knowledge base which could, in principle, incorporate multimedia options, a model of the student, and a set of teaching strategies. The knowledge base is used to impart information directly to students, but it is also used to generate explanations of why students' answers to questions are wrong. This process, in turn, makes use of the model of the student to decide what kinds of misconceptions students will have. Such indirect methods of teaching meet with some success, but they prove comparatively difficult to implement in a tutoring system.

PHILOSOPHICAL ISSUES

AI research, more than that in other sciences, has been surrounded by philosophical controversy. Two related issues have provided the major focus
of debate. The first is whether machines can think, and the second is what role they should be allowed to play in our lives.

The question of whether machines can think, although one that excites the popular imagination, is not necessarily a clear one. One crucial aspect of it, however, is whether there is a difference between computer programs that model phenomena such as the weather, which simulate processes in the world, but do not reproduce them, and AI programs. In other words: is a computer running such a program really intelligent, or is it just simulating intelligent behaviour? On one view, most programs lack real intelligence because they do not interact with the world. The symbols that they manipulate have meaning only because of the way they are interpreted by their programmers. On this view a robot that based its interactions with the world on its internal computations could be intelligent. An opposing view is that real intelligence can be manifest only in biological systems (Searle, 1980). To support this thesis Searle put forward his famous Chinese room argument. If he sat in a room manipulating symbols according to the rules embodied in a computer program, he might, from the outside, be described as reading and answering questions in Chinese. He would not, however, understand Chinese. So, understanding Chinese is not just running a program. However, Searle’s view of what else it is, basically being a biological intelligence, appears to have no foundation, and has been dubbed protoplasm chauvinism (Torrance, 1986).

If machines, or at least robots, can be intelligent, we might at some time in the future have moral responsibilities towards them, or we might be in danger of being dominated by them. To some extent the moral issues raised by such considerations are just those that arise in the application of any science. The difference is that we might be faced not simply with a substance or technique that might be misused, but with something that is itself an “alien” intelligence. However, it is difficult to pinpoint, as Weizenbaum (1976) has tried to do, the sense in which intelligent computers pose a special threat.

ARTIFICIAL INTELLIGENCE, COGNITIVE PSYCHOLOGY, AND THE FUTURE

Since the mid-1970s there has been an enormous growth in AI research. It is no longer possible, as it once was, for an AI researcher, let alone a psychologist, to keep up with developments in all of its subfields. Furthermore, much of AI has become very technical: much more so than cognitive psychology. Nevertheless, the best science often is technical; if cognitive psychologists are not to risk being usurped, they should keep at least one eye on developments in AI.
ARTIFICIAL INTELLIGENCE

FURTHER READING


REFERENCES


INTRODUCTION

In everyday language, "learning" refers to the acquisition of knowledge, but psychologists use the word in a subtly different sense. For psychologists, the acquisition of knowledge belongs to the field of memory, and the word *learning* is usually reserved for changes in behaviour resulting from experience. The word *skill*, on the other hand, has roughly the same meaning in psychology as in everyday usage, but something needs to be said about the distinction between cognitive, motor, and social skills, all of which are covered in this section. Social skills are simply those that are specifically required for effective social interaction. The distinction between cognitive and motor skills is not entirely clear but, roughly speaking, cognitive skills are defined negatively as skills that do not require bodily or perceptual-motor coordination to any significant degree, and motor skills are those that do.

Nicholas J. Mackintosh opens this section by outlining the fundamental principles of classical and operant conditioning in chapter 5.1. Classical conditioning is sometimes called Pavlovian conditioning, after the Nobel Prize-winning Russian physiologist who first investigated it. Operant conditioning is sometimes called instrumental conditioning, because in this type of learning the occurrence of certain elements of behaviour, called responses, are instrumental in eliciting reward or reinforcement. Among the technical terms that Mackintosh introduces in connection with classical conditioning are *unconditional stimulus* (a stimulus that elicits a response unconditionally) and *conditional stimulus* (one that elicits a response only after a process of learning has taken place): there is a potential source of confusion about these concepts that needs to be cleared up. As a result of a mistranslation of Pavlov's writings into English, these are commonly called *unconditioned* and *conditioned* stimuli, and that is also why the learning process that causes a stimulus to elicit a response has come to be called conditioning (a back-formation from "conditioned").

In chapter 5.2 Donald M. Baer introduces applied behaviour analysis,
Psycholinguistics is the study of the mental processes and skills underlying the production and comprehension of language, and of the acquisition of these skills. This chapter will deal with the former aspect only; for the acquisition of language see the suggested “Further reading” at the end of this chapter.

Although the term “psycholinguistics” was brought into vogue during the 1950s, the psychological study of language use is as old as psychology itself. As early as 1879, for instance, Francis Galton published the first study of word associations (Galton, 1879). And the year 1900 saw the appearance of Wilhelm Wundt’s monumental two-volume work Die Sprache. It endeavoured to explain the phylogeny of language in the human mind as an increasingly complex and conscious means of expression in a society, and to describe how language is created time and again in the individual act of speaking. Although Wundt deemed it impossible to study language use experimentally, his contemporaries introduced the experimental study of reading (Huey), of
verbal memory and word association (Ebbinghaus, Marbe, Watt), and of sentence production (Bühler, Seltz). They began measuring vocabulary size (Binet), and started collecting and analysing speech errors (Meringer and Mayer). The study of neurologically induced language impairments acquired particular momentum after Paul Broca and Carl Wernicke discovered the main speech and language supporting areas in the brain’s left hemisphere. In the absence of live brain tomography, aphasiologists began developing neurolinguistic tests for the purpose of localizing brain dysfunctions.

All of these themes persist in modern psycholinguistics. But developments since the 1950s have provided it with two of its most characteristic features, which concern linguistic processing and representation. With respect to processing, psycholinguistics has followed mainstream psychology in that it considers the language user as a complex information processing system. With respect to representation, psycholinguists stress the gigantic amount of linguistic knowledge the language user brings to bear in producing and understanding language. Although the structure of this knowledge is the subject matter of linguistics, it is no less a psychological entity than is language processing itself (Chomsky, 1968). Psycholinguistics studies how linguistic knowledge is exploited in language use, how representations for the form and meaning of words, sentences, and texts are constructed or manipulated by the language user, and how the child acquires such linguistic representations.

I shall first introduce the canonical setting for language use: conversation. Next I shall consider the mental lexicon, the heart of our linguistic knowledge. I shall then move to the processes of speaking and speech understanding respectively. Finally I shall turn to other modes of language use, in particular written language and sign language.

**CONVERSATION**

Our linguistic skills are primarily tuned to the proper conduct of conversation. The innate ability to converse has provided our species with a capacity to share moods, attitudes, and information of almost any kind, to assemble knowledge and skills, to plan coordinated action, to educate its offspring, in short, to create and transmit culture. And all this at a scale that is absolutely unmatched in the animal kingdom. In addition, we converse with ourselves, a kind of autostimulation that makes us more aware of our inclinations, of what we think or intend (Dennett, 1991). Fry (1977) correctly characterized our species as *homo loquens*.

In conversation the interlocutors are involved in negotiating meaning. When we talk, we usually have some kind of communicative intention, and the conversation is felicitous when that intention is recognized by our partner(s) in conversation (Grice, 1968; Sperber & Wilson, 1986). This may take several turns of mutual clarification. Here is an example from Clark and
Wilkes-Gibbs (1986), where subjects had to refer to complex tangram figures:

A: Uh, person putting a shoe on.
B: Putting a shoe on?
A: Uh huh. Facing left. Looks like he's sitting down.
B: Okay.

Here the communicative intention was to establish reference, and that is often a constituting component of a larger communicative goal. Such goals can be to commit the interlocutor or oneself to some course of action, as in requesting and promising, or to inform the interlocutor on some state of affairs, as in asserting, for example. The appropriate linguistic acts for achieving such goals are called speech acts (Austin, 1962).

Although what is said is the means of making the communicative intention recognizable, the relation between the two can be highly indirect. Conversations involve intricate mechanisms of politeness control (Brown & Levinson, 1987). What is conveyed is often quite different from what is said. In most circumstances, for instance, we don’t request by commanding, like in “Open the window”. Rather we do it indirectly by checking whether the interlocutor is able or willing to open the window, like in “Can you open the window for me?” It would, then, be inappropriate for the interlocutor to answer “Yes” without further action. In that case, the response is only to the question (whether he or she is able to open the window), but not to the request.

How does the listener know that there is a request in addition to the question? There is, of course, an enormous amount of shared situational knowledge that will do the work. Grice (1975) has argued that conversations are governed by principles of rationality; Sperber and Wilson (1986) call it the principle of relevance. The interlocutor, for instance, is so obviously able to open the window that the speaker’s intention cannot have been to check that ability. But Clark (1979) found that linguistic factors play a role as well. If the question is phrased idiomatically, involving can and please, subjects interpret it as a request. But the less idiomatic it is (like in “Are you able to...”), the more subjects react to the question instead of to the request.

Another important aspect of conversation is turn-taking. There are rules for the allocation of turns in conversation that ensure everybody’s right to talk, that prevent the simultaneous talk of different parties, and that regulate the proper engaging in and disengaging from conversation (Sacks, Schegloff, & Jefferson, 1974). These rules are mostly followed, and sometimes intentionally violated (as in interrupting the speaker). Turn-taking is subtly controlled by linguistic (especially prosodic) and non-verbal (gaze and body movement) cues (Beattie, 1983).

THE MENTAL LEXICON

Producing or understanding spoken language always involves the use of
words. The mental lexicon is our repository of words, their meanings, their syntax, and their sound forms. A language’s vocabulary is, in principle, unlimited in size. Take, for instance, the numerals in English. They alone form an infinite set of words. But it is unlikely that a word such as twenty-three-thousand-two-hundred-and-seventy-nine is an entry in our mental lexicon. Rather, such a word is constructed by rule when needed. We have the ability to produce new words that are not stored in our mental lexicon.

Figure 1  Fragment of a lexical network. Each word is represented at the conceptual, the syntactic and the sound form level
Source: Bock and Levelt, 1993
How many words are stored? Miller (1991) estimates that the average high school graduate knows about 60,000 words (under one definition of “word”).

One way of representing this enormous body of knowledge is by way of network models. Figure 1 shows a fragment of such a network. Each word is represented by three nodes, one at the conceptual level, one at the syntactic (grammatical) or lemma level, and one at the sound form (phonological) or lexeme level. The lemma is the syntactic representation and the lexeme is the phonological representation. A word’s semantic properties are given by its connections to other nodes at the conceptual level (for instance, that a sheep is an animal, gives milk, etc.). A word’s syntactic properties are represented by its lemma node’s relations to other syntactic nodes (for instance, “sheep” is a noun; French “mouton” has male gender, etc.). The sound form properties, finally, such as a word’s phonological segments, are represented in the way a word’s lexeme node relates to other sound form nodes (“sheep” for instance contains three ordered phonological segments, /ʃ/, /i/, and /p/, as shown in Figure 1).

Different authors have proposed different network models (e.g., Collins & Loftus, 1975; Dell, 1986; Roelofs, 1992), and for different purposes. It is unlikely that such networks can adequately represent all complexities of our semantic, syntactic, and phonological knowledge about words. But they can be useful in predicting speed of word access in comprehension and production, as well as in explaining various kinds of errors that we make in speech production and various disorders of accessing words in aphasic speech.

Especially important for theories of language use are the ways that verbs are represented in the mental lexicon. As a semantic entity, a verb assigns semantic roles to its arguments. The verb walk, for instance, requires an animate argument that specifies the role of agent, as in John walked. The verb greet governs two arguments, one for the agent and one for the recipient of the action, as in Peter greeted the driver. As a syntactic entity, a verb assigns syntactic functions to the sentence constituents it governs. In the above sentence, Peter is the subject and the driver the object. A verb’s argument-function mapping is not random. Most verbs, for instance, map a recipient argument on to a syntactic object function, but not all. The verb receive doesn’t. In Mary received the book, Mary is both recipient and sentence subject. Also, verbs often allow for multiple mappings. In the driver was greeted by Peter, the recipient, not the agent appears in subject position.

For each verb, the mental lexicon contains its possible mapping frames. These play an important role in the speaker’s syntactic planning and in the listener’s syntactic and semantic parsing.

**SPEAKING**

Speaking is our most complex cognitive-motor skill. It involves the conception of an intention, the selection of information whose expression will make
that intention recognizable, the selection of appropriate words, the construction of a syntactic framework, the retrieval of the words’ sound forms, and the computation of an articulatory plan for each word and for the utterance as a whole. It also involves the execution of this plan by more than 100 muscles controlling the flow of air through the vocal tract. Finally, it involves a process of self-monitoring by which speech trouble can be prevented or repaired. The following is a bird eye’s view over these processes.

Conceptual preparation

The question where communicative intentions come from is a psychodynamic question rather than a psycholinguistic one. Speaking is a form of social action, and it is in the context of action that intentions, goals, and subgoals develop. It is not impossible, though, that the intention what to say occasionally arises from spontaneous activity in the speech formulating system itself. It can create rather incoherent “internal speech”, which we can self-perceive. This, in turn, may provide us with tatters of notions that we then consider for expression (cf. Dennett, 1991).

Conveying an intention may involve several steps or “speech acts”. The speaker will have to decide what to express first, what next, and so on. This is called the speaker’s linearization problem (Levelt, 1989). It is especially apparent in the expression of multidimensional information, as in describing one’s apartment (Linde & Labov, 1975). The conceptual preparation of speech, and in particular linearization, require the speaker’s continuing attention. The principles of linearization are such that attentional load is minimized.

Each speech act, be it a request to do $X$, an assertion that $Y$, etc., involves the expression of some conceptual structure, technically called a “message” (Garrett, 1975). That message is to be given linguistic shape; it has to become “formulated”.

Grammatical encoding

A first step in formulating is to retrieve the appropriate words from the mental lexicon and to embed them in the developing syntactic structure. In normal conversation we produce some two words per second. At this rate we manage to access the appropriate words in our huge mental lexicon. Occasional errors of lexical selection (such as “Don’t burn your toes” where fingers was intended) show that the lexicon has a semantic organization.

The standard explanation for such errors is that activation spreads through a semantically organized network, as in Figure 1. In such a network, each node has an activation level between 0 and 1. When the lexical concept node SHEEP is active, then activation spreads to semantically related concept nodes, such as GOAT. Both nodes spread activation “down” to their lemma
nodes. Which one of the lemmas will then be selected for further processing? Normally it will be the most activated one, in this case the lemma for "sheep". But the occurrence of an occasional error shows that there is a small probability that a less activated lemma gets selected. According to one theory (Roelofs, 1992) the probability that a particular lemma becomes selected within a time interval $t$ is the ratio of its activation to the sum of the activation of all other lemma nodes. For instance, if "sheep" and "goat" are the only two active lemmas during interval $t$ after presentation of the picture, and they have activation levels of 0.7 and 0.1 respectively, the probability that the target word "sheep" will be selected during that interval is 7/8, whereas the erroneous word "goat" will be selected with the probability 1/8. Hence, if there is more than one lemma active in the system, there is always a small probability that a non-intended word becomes selected (and it is likely to be semantically related to the target).

Spreading activation theories of lexical selection are typically tested in picture-naming experiments, where naming latencies are measured. For a review of issues in lexical selection, see Levelt (1992a).

As soon as a lemma is retrieved, its syntactic properties become available. Among them are the lemma's grammatical class (preposition, noun, verb, etc.). Each lemma requires its own specific syntactic environment or "frame". Syntactic planning is like solving a set of simultaneous equations. Each lemma's frame has to fit its neighbour's frames, and since Garrett (1975) there are theories about how this is realized (see Levelt, 1989, for a review). Actually, the equations are not quite "simultaneous"; the lemmas for an utterance are typically not concurrently retrieved. Lemmas for salient concepts, such as animate objects, tend to be retrieved faster than for non-salient concepts (Bock & Warren, 1985), and that affects their position in the developing syntactic structure. For a review of grammatical encoding, see Bock and Levelt (1994).

**Phonological encoding**

A selected lemma (but only a selected one: see Levelt et al., 1991) spreads its activation to its lexeme node (cf. Figure 1). At this level two kinds of phonological information become available. The first one is the word's segments, which are "spelled out" one after another. The second one is the word's metrical structure. For "sheep" it is the information that it is a one-syllable word. For "father" it is the information that it is a two-syllabic trochaic word. The metrical frames of successive words are often combined, creating so-called phonological word frames. In *Peter gave him it*, the last three words form one phonological word *gavimit*. In a process of *segment-to-frame association* spelled-out segments are inserted one by one into the corresponding phonological word frames. It is during this ordered insertion that phonological syllables are created, one after another (such as *ga-vi-mit*; see
How this string of phonological syllables determines the precise articulatory gestures to be made by the speech organs is still a matter of much debate (see especially Browman & Goldstein, 1991).

The notion that segments and frames are independently retrieved arose in the analysis of phonological speech errors (Dell, 1986; Shattuck-Hufnagel, 1979). Spoonerisms such as with this wing I thee red, or fool the pill (instead of fill the pool) show that segments can become associated to the right place in the wrong frame.

Phonological encoding also involves the planning of larger units than phonological words. There is, in particular, the planning of intonational phrases. These are units that carry a particular intonational contour. Such contours can be rising, falling or combinations thereof. They often express a speaker’s attitude towards what is said: doubt, certainty, or towards the interlocutor: reassuringness, inviting reaction. See Levelt (1989) for a review of phonological encoding.

The output of phonological encoding is an articulatory programme. Phenomenologically, it appears to the speaker as internal speech. This internal speech need not be articulated. It can be kept in an articulatory buffer, ready to be retrieved for articulatory execution (Sternberg, Wright, Knoll, & Monsell, 1980).

Articulation

The articulatory apparatus consists of three major structures. The respiratory system controls the steady outflow of air from the lungs. The breathing cycle during speech is quite different from normal breathing, with very rapid inhalation and very slow exhalation. The laryngeal system has the vocal cords as its central part. It is the main source of acoustic energy. The vocal tract, finally, contains the cavities of pharynx, mouth, and nose. They are the resonators that filter the acoustic energy in frequency bands or formants. Vowels are characterized by their formant structure. The vocal tract can be constricted at different places, and these constrictions can be made or released in different manners. In this way a wide range of consonantal and other speech sounds can be made.

The control of this utterly complex motor system has been the subject of much research. Present theories converge on the notion of model-referenced control (Arbib, 1981; see also Figure 2). The motor system is given an “articulatory task” (as part of the articulatory programme), such as “close the lips”. There are usually many degrees of freedom in executing such a task. For instance, lip closing can be realized by moving the lips, by moving the jaw, or by doing both to various degrees. The internal model computes the least energy-consuming way of reaching the goal, given the actual state of the articulators (there is continuous proprioceptive feedback to the internal model). The output is a set of efferent control signals to the relevant...

**Self-monitoring**

We can listen to our own overt speech and detect trouble, just as we can listen to the speech of others and detect errors or infelicitous delivery. This involves our normal speech understanding system. We can also detect trouble in our internal speech. When the trouble is disruptive enough for the ongoing conversation, a speaker may decide to interrupt the flow of speech and to make a self-repair.

Not all self-produced trouble (such as errors of selection) is detected by the speaker. Self-monitoring requires attention; we mostly attend to *what* we say, far less to *how* we do it. Detection of trouble is better towards the end of clauses, where less attention for content is required (Levelt, 1989). There are two main classes of trouble that induce repairing. The first one is an all-out error (as in *and above that a horizon*, *no a vertical line*); the error can be lexical, syntactic, or phonological. The second one is that something is not really appropriate (as in *to the right is blue – is a blue point*). The speaker then repairs in order to make the utterance more precise, less ambiguous. Upon detecting either kind of trouble, the speaker can self-interrupt. And this ignores linguistic structure; a speaker can stop in the midst of a phrase, a word, or a syllable. But then, the speaker often marks the kind of trouble
by some editing expression: "no", "sorry", "I mean", for errors; "rather", "that is", for something inappropriate.

Restarting, that is, making the repair proper, is linguistically quite principled. The speaker grafts the repair on to the syntax of the interrupted utterance, which has been kept in abeyance. As a consequence, repairing is like linguistic coordination. One seldom finds a repair such as *is she driving* — *she walking downtown?* And indeed, the corresponding coordination *is she driving or she walking downtown?* is ill-formed. But *is he — she walking downtown?* is a very common repair type, and it corresponds to a well-formed coordination: *is he or she walking downtown?* (Levelt, 1989).

**SPEECH UNDERSTANDING**

The canonical objective in speech understanding is to recognize the speaker's communicative intention. How does the listener induce that intention from the speaker's overt speech, a continuous flow of acoustic events?

Several component processes are involved here. First, there is the hearer's acoustic-phonetic analysis of the speech signal, that is, representing it as a phonetic not just an acoustic event. Second, there is phonological decoding, in particular finding the words that correspond to the phonetic events, and analysing the overall prosodic structure of the utterance. Third, there is grammatical decoding, parsing the utterance as a meaningful syntactic structure. Finally, there is discourse processing, interpreting the utterance in the context of the ongoing discourse, and in particular inferring the speaker's intentions. Let us review these processes in turn.

**Acoustic-phonetic analysis**

It is very hard, if not impossible, to listen to speech as if it were just a string of chirps, buzzes, hums, and claps. We just cannot help perceiving it as speech. In this so-called "speech mode" (Liberman & Mattingly, 1985) we interpret the acoustic event as resulting from a speaker's articulatory gestures as a phonetic event. There is no unanimity in the literature, though, about what kind of representation the listener derives. According to Liberman and Mattingly, the listener derives the speaker's intended articulatory gestures (even if they were sloppy). Others argue that listeners have special detectors for distinctive events in the speech signal, such as for onsets, for spectral peaks, for the frequencies and motions of formants. The detection of such acoustic events may suffice to derive the presence or absence of phonetic features, such as voicing, nasality, vowel height, stridency, and so on (Stevens & Blumstein, 1981).

Speech segments, clusters, and syllables have characteristic distributions of phonetic features. Hence, if such feature detectors are reliable, they may provide sufficient information for effective phonological decoding. Opinions
differ, however, about their reliability. The speech signal is highly variable, dependent as it is on speech rate, sex of the speaker, sloppiness of speech delivery, reverberation or noise in the room, for example. Even if the listener can partial out such effects of the speech context, acoustic-phonetic analysis will often be indeterminate. Still, it may well be sufficient for the purpose. Not every word has to be recognized in order to derive the speaker’s intentions. And where a really critical word is missed, the interlocutor will say “what?” or signal difficulty of understanding in other ways.

For an excellent review of acoustic-phonetic processing, see Pisoni and Luce (1987).

Phonological decoding

Whatever the precise character of the phonetic representations, they are the listener’s access codes to the mental lexicon. How does a listener recognize words in connected speech? A major problem here is to segment the speech, to find out where words begin and end in the continuous flow of speech. There are, basically, two routes here.

The first one is the bottom-up approach, that is, to build on cues in the phonetic representation. Cutler (1990) has argued that English listeners will, by default, segment speech such that there are word boundaries right before stressed syllables. It is a statistical fact of English that 85 per cent of the meaningful words that one encounters while listening begin with a stressed syllable. The segmentation strategy will, therefore, be quite successful. Cutler’s theory has meanwhile found substantial experimental support. Also, there are speech sounds that tend to occur at the ends of words, such as [-ng] and [-nd] for English. Speakers may use such phonotactic properties of their language to predict word boundaries.

The second route is top-down. We often recognize a word before it ends. But that means that we can predict the word’s end, and hence the upcoming word boundary. That gives us a handle on where to start recognizing the subsequent word.

Given that we know a word’s beginning, how do we recognize it? According to the cohort theory (Marslen-Wilson, 1989), a small word-initial feature pattern (corresponding to about two segments of the input word) activates all words in the mental lexicon that match it phonologically. Assume the input word is trespass, and the cluster [tr] has become available. This will activate all words beginning with [tr], such as tremble, trespass, trestle, trombone, etc. This is called the “word-initial cohort”. As more phonetic information becomes available, the cohort is successively reduced. When the vowel [e] is perceived, all items not sharing that vowel, such as trombone, are deactivated. This process continues until a single candidate remains. For trespass this happens when [p] is reached. The segment [p] is, therefore, called the uniqueness point of trespass. A word’s uniqueness point depends
on its word-initial lexical alternatives. For most words the uniqueness point precedes the word’s end.

For an optimally efficient system, the word’s uniqueness point would also be its recognition point. There is good experimental evidence in support of this hypothesis (e.g., Frauenfelder, Segui, & Dijkstra, 1990), though the recognition point may slightly anticipate the uniqueness point in case syntactic or semantic information disambiguates the item from its remaining alternatives (Zwitserlood, 1989). Hence, it will often be possible for a listener to anticipate the upcoming word boundary.

Phonological decoding serves not only the recognition of words, but also their groupings into prosodic constituents, such as phonological and intonational phrases. These constituents carry important information about the syntax of the utterance, and about the communicative intentions of the speaker (cf. Levelt, 1989).

**Grammatical decoding**

As words are successively recognized and prosodically grouped, the listener will as much as possible interpret these materials “on-line” (Marslen-Wilson & Tyler, 1980). Each recognized word makes available its syntactic and semantic properties. There is, then, concurrent syntactic parsing and semantic interpretation, each following its own principles, but interacting where necessary.

In this connection, one should distinguish between local and global syntactic parsing. Local parsing involves the creation of local phrase structure, combining words into noun phrases, verb phrases, etc. There is increasing evidence that local parsing can run on word category information alone (Frazier, 1989; Tyler & Warren, 1987). We have little trouble parsing “jabberwocky” or semantically anomalous prose such as *the beer slept the slow guitar*. Here we construct phrase structure exclusively by recognizing the words’ syntactic categories (Art, Adj, N, V). However, successful local parsing is highly dependent on the intactness of phonological phrases, as Tyler and Warren (1987) could show. For instance, in the above anomalous prose, one should not create a prosodic break between *the* and *slow*, or between *slow* and *guitar*.

Global syntactic parsing, however, interacts with semantic interpretation. In global parsing, semantic roles are assigned to syntactic constituents, and this is to a large extent governed by the verb’s argument/function mapping. When the meaning of words or phrases contradicts the semantic roles they should carry, global parsing is hampered (Tyler & Warren, 1987).

One important aspect of global parsing is the resolution of anaphora. In the sentence *the boxer told the skier that the doctor for the team would blame him for the recent injury*, the anaphor *him* can refer back to *the boxer* and to *the skier*, but global syntax prohibits its referring to *the doctor*. Indeed,
experimental evidence shows reactivation of both boxer and skier, but not of doctor when the pronoun him is perceived. Such reactivation can also be measured for so-called null-anaphors as in the policeman saw the boy that the crowd at the party accused t of the crime. Here there is measurable reactivation of boy at position t (the syntactic “trace” of the boy; see Nicol & Swinney, 1989). But also in this respect global parsing is semantically facilitated, for instance if the anaphor’s referent is a concrete noun (Cloitre & Bever, 1988).

Grammatical decoding doesn’t remove all ambiguity (for instance, the pronoun him above is not fully resolved). Here, further discourse processing is needed.

**Discourse processing**

Partners in conversation construct mental models of the state of affairs they are talking about (Johnson-Laird, 1983; Seuren, 1985). Indefinite expressions (such as in there is a dog in the room) make them introduce a new entity (a dog) in the model. Definite expressions (such as the room in the same sentence) make them look up an already existing entity. The new information in the utterance is then attached to whichever entity it concerns.

Identifying referents is a major accomplishment of human language processing, still unmatched by any computer program. The problem is that referring expressions can be highly indirect. How can a waitress in a restaurant interpret the referent when her colleague says the hamburger wants the bill? Nunberg (1979) argued that there are “referring functions” that map a demonstratum (like the hamburger) on to the intended referent (the person who ordered it). But the range of possible referring functions is almost unlimited. Clark, Schreuder, and Buttrick (1983) and Morrow (1986) have argued (and experimentally shown) that such demonstratum-to-referent mapping depends on the mutual knowledge of the interlocutors and on the saliency of entities in their discourse models.

Indirectness is the hallmark of discourse interpretation. As mentioned above, what is said often relates quite indirectly to what the speaker intends to convey. It is not only politeness that governs such indirectness. All figures of speech, whether polite or not, require the listener to build a bridge from the literal to the intended. This holds equally for metaphor (Sperber & Wilson, 1986), irony (Clark & Gerrig, 1984), and hyperbole (Grice, 1975).

Finally, whereas acoustic-phonetic, phonological, and grammatical decoding are largely automatic processes, discourse processing requires the listener’s full attention. In that respect, it is on a par with the speaker’s conceptual preparation. As interlocutors we are concerned with content. The processing of form largely takes care of itself.
COGNITION

READING

The invention of writing systems, whether logographic, syllabic, or alphabetic, is probably the most revolutionary step in human cultural evolution. It added a powerful means of storing and transmitting information. With the invention of printing, it became a major mechanism for large-scale dissemination of knowledge in a culture.

But equally surprising as this ability to map spoken language on to a visual code is our capacity to efficiently process such a code. When skilled, we silently read five or six printed words per second; this is about twice the rate of conversational speech. This ability has not given us any selective advantage in biological evolution; the invention of writing systems is as recent as about 5,000 years ago. Rather, the ability to read must be due to a happy coincidence of other pre-existing faculties of mind.

One of these is, of course, language. As readers we largely use our parsing potential for spoken language. Visual word recognition feeds into the lemma level of Figure 1. As lemmas are successively activated by the printed words, further syntactic, semantic, and discourse processing operates roughly as for spoken language. There are, admittedly, differences too. There is, for instance, no prosody to help syntactic parsing; instead there is punctuation. Also, there is no external enforcement of rate as there is in speech perception.

Another pre-existing faculty on which reading is parasitic is our enormous ability to scan for small meaningful visual patterns. In a hunter's society these were probably animal silhouettes, footprints, and so on. Words (if not too long or too infrequent) are recognized as wholes; a skilled reader processes a word's letters in parallel. Much ink has been spilled on the question whether the letters individually or the word as a whole activate a phonological code in silent reading, that is, the word's lexeme (see Figure 1). Such phonological recoding indeed exists. But it is only for low-frequent words that this "phonological route" is of any help in lemma access (Jared & Seidenberg, 1991). However, this silent "internal speech" probably does play a role in further syntactic and semantic parsing; it is a way of buffering successive words for further processing.

The ability to scan is optimally used in reading. The basic cycle is this: the reader fixates a word for, on average, one-fifth of a second. The fixation is roughly between the beginning and the middle of the word. During this period lexical access is achieved. In addition, there is some perception of the next word in the periphery of vision. Sometimes this suffices to recognize that next word as well on the same fixation (but the fixation will then last somewhat longer). Usually, however, the information from the periphery of vision is used only to plan a saccadic eye movement (a jump of the eye) to that next word. The size of the saccade depends on the length of the next word; the average saccade is about eight characters in size. The new word is fixated, and the cycle starts all over again.

332
When a word is quite infrequent, or when the reader has trouble integrating it in the developing syntax or semantics, the fixation duration can be substantially longer. Also, the reader may backtrack and refixate an earlier word when there is serious trouble in comprehension.

For a major review of the reading process and its disorders, see Rayner and Pollatsek (1989).

**SIGN LANGUAGE**

Contrary to written language, the sign languages of deaf people are not parasitic on spoken language. They are autonomous languages in the visual mode. Their mere existence shows that our faculty of language is not crucially

---

**Figure 3** Minimal contrasts between signs in American Sign Language: (a) hand configuration, (b) place of articulation, (c) movement

*Source: From Klima and Bellugi, 1979*
dependent on our ability to speak. Deaf children who grow up in a signing deaf community acquire their language at the same age and in roughly the same stages as hearing children do.

Just as words, signs have form and meaning. The articulators of sign language are the hands, the face, and the body. Where words contrast phonemically (for instance in voicing: bath vs path), signs contrast in hand configuration, in place of articulation and in hand movement (see Figure 3). Also, facial features may distinguish between signs.

Although the first coining of a sign is often iconic, its meaning is eventually independent of its form, as it is for words in spoken languages. As a consequence, sign languages are mutually unintelligible, just as spoken languages are (contrary to what Wundt suggested in Die Sprache – see above). Sign languages are rich in morphology (for inflection and for derivation of new signs) and have full-fledged recursive syntax. Many syntactic devices are spatial in character. Anaphora, that is, referring back to an earlier introduced entity, is done by pointing to the locus in the signing space (in front of the body) where the original referent was first “established”. In American Sign Language the sign for transitive verbs either moves from subject to object locus, or from object to subject locus. Each verb has its own “mapping function” (like in spoken language, see above). For the structure and use of British Sign Language, see Kyle and Woll (1985).

There is increasing evidence that a sign language is subserved by the same areas of the brain that sustain spoken language. Poizner, Klima, and Bellugi, (1987) showed that damage to anterior areas of the left hemisphere in native signers resulted in a style of signing highly comparable to the agrammatism of so-called Broca’s patients. Similarly, a form of fluent aphasia resulted when the damage was in a more posterior area of the left hemisphere, comparable to the fluent aphasia of so-called Wernicke’s patients. Damage in the right hemisphere left the signing intact, but patients lost the ability to sign coherently about spatial relations, such as the layout of their apartment. Their spatial representations were damaged, but not their spatial language.

FURTHER READING

REFERENCES


THINKING AND REASONING

Jonathan St B. T. Evans
University of Plymouth, Plymouth, England

The nature of thinking: an historical perspective
Problem solving
Reasoning
Decision making and statistical judgement

Conclusions
Further reading
References

The nature of human thought and the capacity for rational reasoning have been issues of great interest to philosophers and psychologists since the time of Aristotle. Humans have excelled among species in their ability to solve problems and to adapt their environment for their own purposes. We are unique in our possession of a highly sophisticated system of language allowing both representation of complex and abstract concepts and the communication of very precise meaning with one another. We have also developed a new form of evolution — much faster than natural selection — whereby the accumulated knowledge and wisdom of our culture is recorded and passed on through education so that each new generation starts with an advantage on the one before. Despite this impressive record, we also are subject to many systematic errors and biases in our thinking, some of which are discussed in this chapter.

The study of thinking and reasoning in humans can accurately be described as the study of the nature of intelligence. The work described here falls, however, into a quite different tradition from the psychometric study of individual differences in intelligent performance that is usually referred to as the psychology of intelligence. Psychometrics is concerned with the measurement of intelligent performance, whereas the study of thinking and reasoning is
focused on understanding the nature of intelligent processes. Strangely enough, these turn out to be two quite different kinds of undertaking.

THE NATURE OF THINKING: AN HISTORICAL PERSPECTIVE

Historically, we can trace three different conceptions of the nature of thinking. The first of these corresponds to what the non-psychologist might respond if asked to define thought. I shall describe this notion as the contents of consciousness. Common sense (or folk psychology) supposes that we are consciously in control of our actions: we think, therefore we do. When we make a decision or solve a problem it is on the basis of a train of thought of which we are conscious and which we can, if required, describe to another. Such reports of thought are known as introspections. The validity of introspection is clearly assumed in our everyday folk psychology, as we all feel able to ask and answer questions about how and why we have taken particular actions. Indeed, a major industry—opinion polling—is based upon introspectionism. Politicians and political commentators alike are absorbed by the results of polls that ask people not only how they intend to vote, but also to identify the issues which will influence their decisions.

Aristotle and other early philosophers were in no doubt that the mind could and should study itself through introspection. This led to a theory of thinking known as associationism in which thinking was supposed to consist of a sequence of images linked by one of several principles (see Mandler & Mandler, 1964). Associationism and the equation of thought with consciousness remained more or less unchallenged until the late nineteenth and early twentieth centuries when several separate developments conspired to challenge this idea.

First, there were the systematic experimental studies of introspection carried out at the Würzburg School around the beginning of the twentieth century (see Humphrey, 1951). In these experiments, subjects were asked to perform simple cognitive acts such as giving word associations or judging the comparative weight of two objects and then asked to report on what went through their minds at the time. Much to the initial surprise of the researchers, many of these acts did not appear to be mediated by conscious thoughts. Subjects often reported either no conscious experience at all, or else one of indescribable or "imageless" thought.

A second influential development was that of the Freudian school of psychoanalysis which introduced the notion of unconscious thought and motivation. An introspective report of the reason for an action would certainly be suspect to a Freudian since it might well constitute a rationalization of behaviour determined by deep-seated and repressed emotions in the unconscious mind.

The other major influence was the introduction of the school of behaviourism by J. B. Watson (e.g., 1920) whose influence was very strong in
psychology up until the 1950s and which lingers on even in the present day. Watson attacked all study of conscious thought as mentalistic and unscientific. Science, he maintained, could concern itself only with the study of phenomena that were subject to objective observation and independent verification — criteria that introspective reports clearly could not meet. Watson and other behaviourists effectively redefined thought as simply complex forms of behaviour which were the result of stimulus–response learning. Study of stimulus–response pairings and reinforcement history were sufficient to explain all phenomena attributed — by the mentalistically inclined — to thinking.

From the viewpoint of a modern cognitive psychologist both introspectionists and behaviourists might be seen as half right. The behaviourists were probably right in their contention that thought cannot be studied effectively via introspection. The mentalists, on the other hand, were correct in asserting that complex behaviour could not be explained without reference to internal mental processes. Their mistake — with the benefit of hindsight — was to assume that such processes were necessarily conscious and reportable. This leads us to the third conception of human thought — that of information processing.

Psychologists' own thinking — like that of their subjects — is constrained by the availability of models and analogies. Watson used the analogy of a telephone exchange to explain his notion of learning by stimulus–response connections. Although its origin can be traced to earlier, highly creative thinkers (especially Craik, 1943) the emergence of cognitive psychology in the 1950s and 1960s was largely due to the development of cybernetic systems and then the digital computer. Computers are general-purpose information processing systems. They compute by manipulating symbols which can represent almost anything — numbers and arithmetical operators, permitting arithmetic; letters and words as in word and text processing; collections of facts stored in a database; and so on.

When people perform mental arithmetic, we would describe this as an act of thought. So is a computer also thinking when it performs computations to solve problems? It appears that it is, although some philosophers (e.g., Searle, 1980) maintain that computer intelligence is intrinsically different from that of the human mind. The point of the analogy, however, is that we can see that computers can perform complex acts of information processing — depending upon their programming — but without any need to assume that they are conscious. Once you equate thinking with information processing, then the task of the modern cognitive psychologist is clear: understanding thought is the problem of discovering the software of the human brain. Many psychological theories in fact are formulated as working computer programs which attempt to simulate the behaviour of a human being who is solving a problem or engaged in some other cognitive activity.

In spite of this advance, arguments persist among cognitive and social
psychologists as to the value of introspective reports. Some cognitive psychologists disregard them entirely on the basis of much evidence that such reports can be both incomplete and misleading (Nisbett & Wilson, 1977). One interesting line of argument is that verbal reports are useful indicators of thought processes but not as used in the tradition of introspective reporting (Ericsson & Simon, 1980). According to this view, verbalizations are the products of cognitive processes and can be fruitfully interpreted by the psychologist when subjects are asked to "think aloud" while performing a task or solving a problem. Introspective reports fail because first, they are retrospective rather than concurrent, and second, they invite subjects to describe their thinking or to theorize about the causes of their behaviour.

The psychology of thinking can be broadly defined to cover a wide range of topics. For example, Gilhooly (1982) distinguishes between directed thinking - as found in problem solving and reasoning - undirected thinking - as in day-dreaming - and creative thinking. In this chapter we shall focus on directed thinking: thought aimed at achieving specific goals. This is an area in which reasonable theoretical progress has been made, and for which there are clear practical applications in everyday life.

Studies of directed thinking fall broadly into three main areas which are described as problem solving, reasoning, and decision-making. We shall consider each in turn.

**PROBLEM SOLVING**

A person has a problem whenever he or she wishes to achieve a goal and is unable to proceed immediately to do so. Problem solving consists of finding a method of getting from where you are to where you want to be, using such resources and knowledge as you have available. This definition obviously covers a vast range of human activity; problem solving is clearly involved in solving crossword puzzles and choosing chess moves, but it is equally involved in finding your way to a new destination, obtaining a ticket for a sold-out sporting contest, or working out how to persuade your boss to give you a pay rise.

One distinction which has helped psychologists think about the vast range of behaviours involved in problem solving is that between well-defined and ill-defined problems. In a well-defined problem, all the information needed and the means of solution are available at the outset. This is typical of things that are set as "problems" in newspapers, and so on, and also typical of much research in the psychological laboratory. An anagram is an example of a well-defined problem. You know the letters that constitute the solution word and also the means of solving the problem - rearrangement of the order of letters - at the outset. Well-defined problem solving thus consists of applying known rules to known information in order to transform the situation and achieve the goal.
Some of the most famous studies of well-defined problem solving were conducted by Newell and Simon (1972). An example of one of their problems is cryptarithmetic, in which subjects were given the following problem:

\[
\begin{array}{c}
\text{DONALD} \\
+ \text{GERALD} \\
\hline
\text{ROBERT}
\end{array}
\]

Subjects are also told that D = 5 and that each letter represents a single digit number between 0 and 9. Given this information and the assumption that the normal rules of arithmetic apply, it is possible — though complicated — to work out what all the letter-number pairings must be. If the reader wishes to attempt this problem, then it is suggested that a good record (on paper) of the sequence of attempts — including errors and correction — be kept.

Newell and Simon (1972) made an important theoretical contribution with the idea of problem solving as a search through a problem space. A problem space consists of a number of linked states including an initial or starting state and one or more goal states. All problems include permissible operators which allows one state to be transformed to another. Thus, solving problem consists in applying operators repeatedly to transform the initial state into a goal state.

As an example consider the game of chess (also studied by Newell & Simon, 1972). The states of the game can be described as the position of the pieces on the board plus some additional information (whose turn is it to move, do players have the right to castle, may a pawn be captured \textit{en passant}, and so on). The initial state is thus the board with the pieces in starting position with White having the right to move. A goal state is any position in which the player has won the game either by checkmating the opponent or making such a mate inevitable. The permissible operators are the laws of chess, which determine the moves that can legally be made in a given situation.

Note that these definitions tell us nothing about the strategy of chess. The problem space consists of all states that can be reached by legal moves — a vast number of possibilities in the case of chess. The strategy of the game obviously consists in choosing between alternative legal moves in such a way as to move towards the goal state of a winning position. In chess, as in many other problems, the problem space is too large for an exhaustive search to be feasible. You cannot consider all moves and all possible replies to more than a very few moves ahead without the number of possible positions becoming enormous. Thus Newell and Simon (1972) emphasize the importance of heuristic strategies. An heuristic is a short-cut, rule of thumb method which may lead to a quick solution, but which may also fail. What heuristics do is to drastically reduce the size of the problem space to be searched in the hope that the goal state is not excluded in the process.
Consider the following anagram: GBANRIEK. Since it has eight letters the total problem space includes the $8! = 40,320$ possible rearrangements of the letters. A guaranteed, algorithmic (i.e., exhaustive search) method of solving this involves constructing all 40,320 letter strings and checking whether each is a word. A typical heuristic method, on the other hand, might involve looking for familiar letter patterns to decompose the problem. For example, we note that the anagram includes the letters I, N, and G and speculate that the word might be of the form ____ING. Thus we have now reduced the problem to solving the five-letter anagram BAREK which has only 5! (120) possible solutions and is thus much easier. We may now spot the solution word BREAKING. Like all heuristics, however, this was not guaranteed to work. Many words contain the letters ING in other configurations, e.g., GELATIN.

Problem space analysis is extremely useful as it provides a common framework in which to describe a very wide range of different problems. Newell and Simon (1972) studied subjects using think-aloud protocols while solving problems such as the cryptarithmetic example given above. They concluded that people have sets of general-purpose problem solving strategies that are used in similar ways to search problem spaces, no matter what particular domain is involved. They implemented their theory in a working computer program called General Problem Solver that was claimed to solve the same problems as the human subjects and in a similar way.

Important though this work has been, the conclusions are somewhat questionable. The first difficulty is that most real-life problems are ill defined. Some aspect of the problem — the information assumed, the means of solution, sometimes even the goal — is incomplete or missing at the outset. Take the case of engineering design which was subjected to detailed psychological study by Ball, Evans and Dennis (in press). An engineer is given a general specification for a device which includes its functionality — what it must do — and a number of constraints, including costs. The engineer must then come up with a technical specification for a device which can be constructed and can be demonstrated to work.

As Ball discovered, such problems are not at all well defined. Nearly all the information required to solve the problem is implicit and must be retrieved either from the existing knowledge and experience of the engineer or by researching technical manuals, and so on. In the process of design, constraints emerge that were not apparent at the outset. The goal initially set may also be modified and rethought as the work progresses. Now such activity can still be usefully described within the problem space framework — a space that is being continually augmented and redefined by the knowledge and experience of the engineer. However, the point is that simply applying the problem space description provides no explanation for some of the most important aspects of the process, particularly the means by which prior knowledge and experience are retrieved and applied.
A number of more recent studies of human problem solving have focused on ill-defined problems and the use of prior knowledge. Of particular interest has been the role of analogy in solving problems (see Gick & Holyoak, 1980, 1983; Keane, 1988). Most real-life problem solving — including “expert” problem solving — occurs within contexts where the solver has previous experience. Clearly, people do not solve all such problems as if seen for the first time; they must extrapolate from past experience. The theoretical and practical interest lies in how they actually bring their prior knowledge to bear.

A problem that has featured in many of these studies is the **tumour problem** first introduced by the Gestalt psychologist Duncker (1945). The problem is that of a patient who has a malignant but inoperable tumour that can be destroyed only by radiation. However, the radiation destroys healthy tissue at the same rate as diseased tissue. The solution that subjects must find is to use a lens to converge the rays at the point of the tumour. Hence, the rays accumulate only to sufficient intensity to destroy the tumour and not the healthy tissue they pass through on the way (see Figure 1).

The problem is incompletely defined in that while the goal and constraints are generally indicated, subjects must search their knowledge and imagination for possible means of solution. General knowledge of medical procedures is unhelpful; surgery is out by definition; drug treatments are of no relevance. The problem can, however, be facilitated by provision of a structural analogue such as the General story. The General is trying to attack a fortress which is well defended and which may be reached by a number of different roads. Each road is mined and may be safely crossed only by a small band of men. The General splits his force into small groups which approach

![Figure 1 Solution to Duncker's tumour problem](image)
THINKING AND REASONING

simultaneously from different directions, and converge at the fortress with sufficient force to win the battle.

Gick and Holyoak (1980) showed that presentation of the General story could facilitate convergence solutions to the tumour problem provided that subjects were given a cue as to its relevance. There is a theoretical argument as to whether analogies can work by direct mapping of the elements of the analogy on to the problem, or whether the solution is mediated by an abstract schema. Gick and Holyoak suggest that subjects may construct and apply a convergence schema which is defined in terms of variables. For example, in the schema the goal is to destroy an obstacle, the means is a sufficient force, the constraint is that direct application is blocked, and so on. The General story could lead to development of a schema which is applied to the tumour problem.

The notion of schema is a useful one, in that it helps us to understand how knowledge may be abstracted, generalized, and applied in new situations. The notion will recur in the discussion of reasoning to which we now turn.

REASONING

Reasoning is the process of drawing conclusions or inferences from given information. An important distinction is that between deductive and inductive inference. Deductive reasoning involves drawing conclusions that are logically valid, that is, they necessarily follow from the premises on which they are based. Thus such inferences do not increase the amount of information contained in the premises; they merely render explicit what was previously latent information. The following are examples of valid deductive inferences:

The television will work only if it is plugged into the mains;
The television is not plugged into the mains,
Therefore, the television will not work.

John is taller than Jim;
Paul is shorter than Jim,
Therefore, John is taller than Paul.

The validity of the first example does not depend in any way on our knowledge of television sets, but only on our understanding of the connective “only if”. Any argument of the form \( p \text{ only if } q; \not q, \therefore \not p \) would be logically valid no matter what propositions we substitute for \( p \) and \( q \). Hence, validity depends on the form of the argument, not its actual content. In logic, the statement \( p \text{ only if } q \) cannot be true in a world where \( p \) is the case and \( q \) is not the case. Hence, once we know that \( q \) is false we can infer that \( p \) must be false as well.

The second example requires us to know that the relation taller—shorter is transitive. A transitive relation is one where the objects are ordered in a single
line so that whenever \( A \) is higher than \( B \) on the scale, and \( B \) is above \( C \) then \( A \) is also above \( C \). Examples of other transitive relations are better–worse, warmer–colder, and darker–lighter. Many relations, of course, are not transitive. If \( A \) is next to \( B \) and \( B \) is next to \( C \) it does not follow that \( A \) is next to \( C \).

Deductive inferences are very important in intelligent thinking as they allow knowledge to be stored in generalities and then applied to particular situations. Thus if we want to watch television and discover one that is unplugged, we immediately plug it in. This is a simple example of reasoning in order to solve a problem. The limitation of deductive reasoning, however, is that it adds no new knowledge; thus we cannot learn by deduction. Induction is involved whenever our conclusion has more information than the premises. A typical example is an inductive generalization such as

The Australian soap operas I have seen were boring, hence all Australian soap operas are boring.

Such an inference is clearly not logically valid, though it could well influence what you watch when you get the TV plugged in.

The British psychologist, Peter Wason, invented two famous problems that have been used extensively to study both inductive and deductive reasoning. The inductive problem was first published by Wason (1960) and is known as the “2 4 6” task. The subjects are told that the experimenter has a rule in mind which applies to “triples” of three whole numbers. An example which conforms to the rule is “2 4 6”. The subjects are then asked to discover the rule by generating triples of their own. In each case the experimenter says whether the triple conforms or not. Subjects are told to announce the rule only when they are very sure that they know it.

The actual rule is “any ascending sequence” but the subject is induced by the example to form a more specific hypothesis, such as “ascending with equal intervals”. Most subjects have great difficulty in solving the problem initially because all the examples they test appear to conform to the rule. The reason is that subjects test positive examples of their hypothesis which invariably turn out to be positive examples of the experimenter’s rule as well. Their hypothesis can be refuted only by testing a negative example of the hypothesis such as “1 2 4” which is revealed as a positive instance of the actual rule. The set relationships involved are shown in Figure 2.

The protocols discussed by Wason (1960) were very interesting, suggesting that some subjects became so convinced of the correctness of their hypotheses that they were led to reformulate the proposed rule in different terms when told it was wrong. A striking example of this is shown in Table 1.

Wason’s interpretation of his findings was that subjects have a confirmation bias, meaning that they systematically seek out evidence that confirms rather than refutes their current hypothesis. He suggested that such a confirmation bias is a very general tendency in human thought which may
Table 1  Example protocol from Wason (1960)

8 10 12: two added each time; 14 16 18: even numbers in order or magnitude; 20 22 24: same reason; 1 3 5: two added to preceding number.

The rule is that by starting with any number two is added each time to form the next number.

2 6 10: middle number is arithmetic mean of other two; 1 50 99: same reason

The rule is that the middle number is the arithmetic mean of the outer two.

3 10 17: same number, seven, added each time; 0 3 6: three added each time.

The rule is that the difference between two numbers next to each other is the same.

12 8 4: the same number subtracted each time to form the next number.

The rule is adding a number, always the same one, to form the next number.

1 4 9: any three numbers in order of magnitude.

The rule is any three numbers in order of magnitude.

(17 minutes)
account for the maintenance of prejudice and false belief. While a number of authors have accepted this interpretation, it has also been subject to serious challenge (see Evans, 1989; Klayman & Ha, 1987).

The problem is that the subjects in the “2 4 6” experiment have no way of knowing that a positive test cannot lead to refutation of their hypothesis, and in many real-world situations it would do so. For example, in science it is customary to formulate general hypotheses and test if they apply to specific cases. Hence, given the hypothesis “All metals expand when heated” you would test any untried metal to see if the prediction holds – and if it did not you would indeed refute the hypothesis. You would not be likely to try heating non-metal things, and even if you did and they expanded, it would mean only that your rule was insufficiently general.

Arguments such as these have led some authors to suggest that subjects’ behaviour on the “2 4 6” is more rational than it at first appears and that if there is a bias, it is towards positive testing rather than to confirmation as such. A particularly interesting experiment reported by Tweney, Doherty, and Mynatt (1980) provides evidence for this. In one study, instead of defining instances in positive and negative terms (right/wrong, belonging/not-belonging) they told subjects that all triples were either MEDs or DAXes and that “2 4 6” was an example of a MED. What happened was that subjects continued to test their hypotheses positively but alternated between testing MED and DAX hypotheses. For example, if the hypothesis was that “triples ascending in equal intervals are MEDs and others are DAXes”, then they might test “1 2 5” predicting it to be a DAX. This meant that they effectively tested negative examples of the usual hypothesis and hence solved the problem much more easily. The psychological difference is that the negative test of MED was construed as a positive test of DAX.

A close parallel to these findings occurs with the second and most famous of Wason’s problems – the four-card selection task (see Evans, Newstead and Byrne 1993 for detailed review and discussion). This problem requires subjects to test hypotheses via deductive reasoning. In the classic “abstract” version of the task, subjects are told that a set of cards always has a capital letter on one side and a single-figure number on the other side. They are then shown four such cards lying on a table with the exposed values as shown in Figure 3. The subjects are told that the following rule may be true or false:

*If there is an A on one side of the card then there is a 3 on the other side of the card.*

The subjects’ task is to turn over those cards – and only those cards – that are needed to decide whether the rule is true or false. The task is deceptively simple, since most subjects fail to solve it. The common answers given are A alone, or A and 3. The correct answer is the A and the 7. The reason is that the rule can be shown to be false only if there is an A on one side of a card and number other than a 3 on the other. Only by turning the A and the 7 (not a 3) is it possible to discover such a card. There is also no point
Wason's original claim was again that card selections reflected a confirmation bias: subjects were trying to prove the rule true rather than false, that is, looking for the combination A and 3, rather than A and not-3. This view was, however, refuted to the satisfaction of Wason as well as other authors by the finding of "matching bias" reported by Evans and Lynch (1973). They pointed out that the preferred selections, A and 3, were not only the verifying choices, but also the positive choices matching the items named in the actual rule. Verification and matching could, however, be separated by introducing negative components into the rule. Consider for example, the rule:

*If there is an A on one side of the card then there is NOT a 3 on the other side of the card*

If subjects have a confirmation bias, then they should now choose the A and the 7 which confirm the two parts of the rule. If, however, they have a matching bias then they should continue to choose A and 3 which are the correct and falsifying combination on this rule. Subjects do, in fact, continue to choose predominantly matching values on this and other variants of the rule, thus confirming the predictions of Evans and Lynch. Evans (1989) regards matching as an example of a generalized positivity bias, that is, bias to think about positively defined items, which also accounts for subjects' behaviour on the "2 4 6" task.

Dozens of experiments have been published – and continue to be published – in which subjects are asked to solve versions of the Wason selection task. Most of these have been concerned with the so-called thematic materials facilitation effect. This has its origin in two early studies discussed in Wason and Johnson-Laird's (1972) famous textbook on reasoning. In one of these (Johnson-Laird, Legrenzi, & Legrenzi, 1972) subjects were shown envelopes in place of cards, together with the following Postal Rule:

*If the letter is sealed then it has a 50 lire stamp on it.*

Subjects were then shown four envelopes which were either front side up and showing a 50 or 40 lire stamp, or rear side up showing that they were sealed or unsealed (see Figure 4). The subjects had to decide which envelopes to turn
Figure 4 The four envelopes shown to subjects in the Postal Rule version of the selection task

over in order to decide if the rule was true or false. The usual matching response on the abstract task would lead to choice of the sealed envelope and the 50 lire stamp. However, almost all subjects made the logically correct choice of the sealed envelope and the one showing a 40 lire stamp.

The original interpretation offered of this and other similar experiments was that use of thematic materials facilitated logical reasoning on the task. This view has been considerably refined by subsequent research, however. The problem with the Postal Rule is that a very similar rule (involving pence rather than lire) was in force in the UK at the time of the study. Thus it was argued that subjects knew from experience that envelopes with a lower value stamp must not be sealed and that hence no “reasoning” as such was required to solve the problem. This argument was supported by the findings of several later studies which showed that first, the Postal Rule produces no facilitation of performance in American subjects unfamiliar with such a rule, and second, British subjects too young to remember the rule (it was dropped in the 1970s) show no facilitation on the problem whereas older subjects perform much better.

It is not the case, however, that subjects must have direct experience of the context in order for a problem content to facilitate on the selection task. A very effective version, for example, is the Sears Problem in which subjects
are asked to play the role of a store manager checking that a company rule has been followed. The rule is

*If a purchase exceeds $30, then the receipt must be approved by the departmental manager.*

Subjects are shown four receipts, two of which are front side up showing totals of above and below $30 and two of which are front side down and either have or do not have the signature of the departmental manager on them. Few subjects have any difficulty in correctly deciding to turn over the receipt for more than $30, and the one that has *not* been signed by the manager. This is despite the fact that subjects have not worked as managers in department stores.

While arguments exist about the precise reason for facilitation of performance by these kinds of thematic content, the general idea is that where subjects have either direct or analogous experience that can be linked to the problem, then they can solve it. Another line of argument is that it is the introduction of deontic terms such as *may* and *must* which carry with them notions of permission and obligation that causes the facilitation. The idea is that we have generalized reasoning schemas that enable us to understand the logic of any situation in which, for example, a precondition is set for an action. Thus, once we have identified the action (e.g., sealing an envelope, spending over $30) and the precondition (sufficient value stamp, permission of departmental manager) we know what to do: we are applying a generalized permission schema to the problem at hand.

The two problems of Peter Wason discussed in this section have stimulated much interesting psychological work on the nature of human reasoning. The specific findings discussed here invite two general conclusions: first, that reasoning with “abstract” problem material is heavily biased by a tendency to think about positively rather than negatively defined information, and second, that the introduction of thematic problem content, and hence associated prior knowledge, can have a dramatic effect on the reasoning observed, and sometimes produces much better logical performance. The “sometimes” in the latter conclusion is needed. Other research, which there is no space to discuss here, has also indicated that prior knowledge can be a source of bias and error in reasoning. This is especially the case when subjects are asked to evaluate the logic of an argument but have strong prior beliefs about the truth of a conclusion (see Evans, 1989, chap. 4).

**DECISION MAKING AND STATISTICAL JUDGEMENT**

In a problem solving task, it is normally possible to work out and demonstrate a solution to the problem set. Once you have the solution, you know it and can prove it. In a decision-making task, however, subjects are required to exercise judgement about a choice that will only later prove to work out
well or badly. Decision-making means committing yourself to choices between actions by anticipation of what the outcomes will – or may – be. Thus when we make any decision – to accept one job rather than another, to marry someone or not, go to a football match rather than stay at home – we do so in the hope that the future we chose was to be preferred to the one we avoided.

Decision-making is obviously of great importance in the real world, but it is a subject of considerable psychological interest too. Most real-world decision-making is done under conditions of uncertainty: we do not know for sure what will happen as a result of each choice and at best can try to estimate the probabilities of different outcomes. If we are to choose rationally then we need to evaluate the desirability of these outcomes as well. In the parlance of decision theory, we should try to maximize expected utility where utility is the subjective value of the outcome and where the term “expectation” means that we weight the various possible outcomes by their likelihood of occurring. Hence, a small chance of a highly desirable outcome might be equally attractive to a much better prospect of a less desirable outcome.

There has been much debate in the psychological literature about whether people choose rationally or not. The notion of rational choice has several components. First, it implies that people will consciously consider the various actions available to them and try to project ahead the possible outcomes and further choices to which they lead in what is termed a decision tree. Second, it is assumed that they assign probabilities and utilities to each of these outcomes as accurately as possible in the light of their current beliefs. Finally, rational decision-makers are assumed to apply systematic principles, such as the maximization of expected utility, in order to decide their final choices.

There are many demonstrations of human choice behaviour that appear to depart from this idealized notion. Within the space restriction here I shall discuss just one aspect – the ability of people to judge probabilities or to reason statistically. A famous set of papers by the psychologists Amos Tversky and Daniel Kahneman dating from the early 1970s have apparently demonstrated the frailty of human probability judgement. This research is often cited as evidence of irrationality, although Tversky and Kahneman themselves follow the tradition of work on “bounded rationality” espoused by Newell and Simon (1972). The idea is that people cannot base their probability judgements on probability theory due its computational complexity and instead employ short-cut rules of thumb known as heuristics. While often useful, such heuristics can also lead to systematic errors and biases.

Of the heuristics discussed by Kahneman and Tversky, the two most famous are those of representativeness and availability (see Kahneman, Slovic, & Tversky, 1982 for a collection of relevant papers, including the seminal ones). Probability or frequency of an event is estimated by the availability heuristic when people base their judgement on the ease with which examples can be brought to mind. Such a heuristic would often be effective.
For example, an experienced doctor might base a provisional diagnosis on her recollection of the numbers of previous cases or patients with similar symptoms who turned out to suffer from a particular condition. Assuming that memory was accurate and that experience was representative then this is a good, if rough basis for a judgement.

As Tversky and Kahneman have demonstrated, however, relying on availability of recalled examples can lead to biases. For example, some types of information are easier to retrieve than others, due to the way in which memory is organized. For example, most people will say, if asked, that there are more words in English that start with the letter \( k \) than those that have \( k \) as the third letter, although the reverse is true. The problem is that it is hard to generate examples of the latter category: they cannot easily be “brought to mind”.

Availability is also implicated in biases which preserve false beliefs and theories. An interesting example is the phenomenon of illusory correlation. It has been demonstrated in a number of studies that human judges — including experts — hold theories that are not supported by the evidence they encounter. For example, some clinicians maintain that projective personality tests such as the Rorschach ink blot test is useful in diagnosing mental illness despite a lack of any supporting evidence. Research has shown that such judges perceive a correlation between test results and diagnoses in a set of data in which they are in fact randomly related. A plausible explanation of illusory correlation is that the judges selectively remember the cases that confirm their expectations or pet theories. Thus confirming cases are more available in later recall and bias the judgement of the correlation.

The representativeness heuristic is involved in judgements of conditional probability. The likelihood of a sample given a population, or of an event given a hypothesis is dependent upon the perceived similarity of the two. Similarity judgements may, however, cause the subject to overlook the relevance of a critical statistical feature such as the size of the sample, or the base rate occurrence of the event. A simple example is provided by the conjunction fallacy (Tversky & Kahneman, 1983). Subjects are given a description of Bill as follows:

Bill is . . .

They are then asked to rank the likelihood of several statements including the following:

\( a \) Bill is an accountant
\( b \) Bill plays jazz for a hobby
\( c \) Bill is an accountant who plays jazz for a hobby.

What happens is that most subjects rate the order of likelihood of these statements as \( a > c > b \). However, there is a statistical impossibility here in that statement \( c \) cannot be more likely than statement \( b \). Given two events \( A \) and
The probability of them both occurring – $P(A \cap B)$ – must be less than or equal to the probability of either $P(A)$ or $P(B)$. Whenever $c$ is true then $b$ is true as well, because Bill plays jazz for a hobby. If all jazz players were accountants then the two statements would be equally likely, otherwise $b$ has to be more probable.

The explanation offered for the fallacy is that the description of Bill conforms to our stereotype for accountants but not for jazz players. Thus the statement $c$ is more representative of the description than is statement $b$ and hence judged more probable.

One of the most famous of Kahneman and Tversky’s problems is the Cabs Problem. You are given the following information: in a certain city there are two cab companies: the Blue cab company, which has 85 per cent of the city’s cabs, and the Green cab company, which has 15 per cent of the city’s cabs. A cab is involved in a hit-and-run accident and a witness later identified the cab as a Green one. Under tests the witness was shown to be able to identify the colour of a cab correctly about 80 per cent of the time under comparable viewing conditions. The subjects are asked if the cab involved in the accident is more likely to have been Green or Blue. Most say Green, although the correct answer is Blue.

The problem is that subjects disregard the base rate or prior probability of the cab colour – 85 : 15 in favour of Blue. In fact, when asked to give a numerical estimate, most subjects say 80 per cent Green – the chance of the witness correctly identifying a cab. If there were no witnesses, it would be obvious that the chance of the cab being Blue was 85 per cent – the base rate. As Figure 5 shows, however, the chance of a Blue cab being identified as Green is 17 per cent which is still higher than the chance (12 per cent) of a Green cab being identified as Green.

$$A \text{ Probability of Blue identified as Blue } = 80\% \times 85\% = 68\%$$

$$B \text{ Probability of Blue identified as Green } = 20\% \times 85\% = 17\%$$

$$C \text{ Probability of Green identified as Green } = 80\% \times 15\% = 12\%$$

$$D \text{ Probability of Green identified as Blue } = 20\% \times 15\% = 3\%$$

*Figure 5* Probabilities in the Cabs Problem

354
Originally, the base rate fallacy was interpreted as the base rate lacking representativeness, although the explanation is probably more fundamental. We find it very difficult to apply abstract statistics to individual cases. Hence, many cigarette smokers are aware of the statistical risks for smokers as a whole, but do not feel that this affects them as individuals. However, we can apply statistics when we see a causal connection. If the cabs problem is slightly reworded, most subjects give the right answer. In this version the number of Green and Blue cabs in the city is the same, but 85 per cent of the cabs involved in accidents are Blue. The image of reckless Blue cab drivers conjured up induces subjects to take account of the base rate, although from a statistical point of view the problem is unchanged.

CONCLUSIONS

Psychological research on thinking and reasoning has produced some useful — and sometimes surprising — conclusions. The common-sense view, that intelligent actions are based on conscious and rational acts of thinking, does not fit the evidence at all well. If thought is to be defined as the information processing that underlies problem solving, reasoning, and decision-making, then surprisingly little of this appears to be accessible through introspection.

If human thinking is rational — and the success of the species suggests that it should be — then that rationality is highly constrained by our capacity to process information. In particular, we seem to solve problems and make decisions largely on the basis of heuristic processes which serve us well in some circumstances, but lead us into error and bias in others. We seem to have particular difficulty in understanding probability and uncertainty despite the crucial role that this plays in rational decision-making.

Studies of reasoning also show that we are prone to biases, for example in a strong preference for thinking about positively defined information. Perhaps the most important finding in this area, however, is the discovery that we do not — as was once thought — appear to reason by the use of an abstract mental logic, but instead seem to be highly influenced by the content and context of the problems with which we are faced. The processes of human thought appear to be quite specific to the areas of knowledge which we are involved in applying.

FURTHER READING


**REFERENCES**


356


ARTIFICIAL INTELLIGENCE

Alan Garnham
University of Sussex, England

<table>
<thead>
<tr>
<th>History</th>
<th>Applications</th>
</tr>
</thead>
<tbody>
<tr>
<td>Knowledge representation</td>
<td>Philosophical issues</td>
</tr>
<tr>
<td>Vision</td>
<td>Artificial intelligence, cognitive</td>
</tr>
<tr>
<td>Thinking, reasoning, problem solving</td>
<td>psychology, and the future</td>
</tr>
<tr>
<td>Language</td>
<td>Further reading</td>
</tr>
<tr>
<td>Learning</td>
<td>References</td>
</tr>
</tbody>
</table>

Artificial intelligence, almost always known as AI, attempts to understand intelligent behaviour, in the broadest sense of that term, by getting computers to reproduce it. “Intelligent behaviour” is taken to include thinking, reasoning, and learning, and their prerequisites (perception, the mental representation of information, and the ability to use language). Indeed, much current work in AI is concerned with modelling aspects of behaviour that would not normally be thought of as requiring any special intelligence. As part of computer science, AI is separate from cognitive psychology, although there is a large overlap in subject area. The two come together (with, most importantly, linguistics and philosophy) in the multidisciplinary approach of cognitive science.

Although AI aims to understand human intelligence, it also aims to produce machines that behave intelligently, no matter what their underlying mechanism. However, although these machines may not model human behaviour, their construction may reflect principles that are useful in studying it.
Since AI depends on computers, it is a relatively new discipline: the name was first used in the mid-1950s, though a few years earlier, pioneers such as Alan Turing in Britain and Claude Shannon in the United States had worked out how to write chess-playing computer programs. The dream of mechanized thought has, of course, a much longer history. The philosophers Blaise Pascal (1623–1662) and Gottfried Leibniz (1646–1716) built small calculating machines, and conceived grander schemes for formalizing thought processes. Charles Babbage (1792–1871) came nearer to building a universal computing machine, but was foiled by the limitations of having to use mechanical parts. Real computers had to wait for electronic components — first vacuum tubes, then semiconductors.

A conference at Dartford College, New Hampshire, in 1956 effectively launched AI research, even though its organizers felt disappointed at the time. In retrospect, the most important line of research discussed at the conference was that of Allen Newell and Herbert Simon (see e.g., Newell, Shaw, & Simon, 1957) on human problem solving. They proposed the idea of a heuristic (“rule-of-thumb”) procedure for solving problems, and they shunned a line of research based on modelling the properties of networks of brain cells, which only assumed major importance again 25 years later, in the guise of connectionism. Newell and Simon's information processing approach was the dominant one in the early days of AI, and it remained influential throughout the 1960s — the so-called semantic information processing era.

There was, however, a subtle shift of emphasis from a formal analysis of tasks to one based on the meaning of the information being processed. Furthermore, in attempting to tackle broader problems, such as natural language understanding, AI researchers quickly discovered that everyday tasks depend on huge amounts of background knowledge. To keep programs manageable, they were made to work in limited domains, in particular BLOCKSWORLD — a tabletop with prismatic blocks on it. It was hoped that programs that worked in these limited domains would scale up to real situations. In practice they did not, and in retrospect it is often obvious why they could not.

The 1970s was a somewhat disappointing period in "traditional" areas of AI research. Indeed, in the UK the Lighthill report (Lighthill, 1972) concluded that AI should not be a priority area for research. The late 1970s saw four important developments. The first was a shift in interest from specific computer programs to general principles. To some extent this development was linked to the second, the emergence of cognitive science, in which AI techniques are used with the primary goal of developing general theories of cognition, rather than with the more applied ("engineering") goal of building intelligent machines. The third development was a shift in the research topics seen as central to AI. In particular, fifteen years of research on the first expert systems was beginning to have spectacular payoffs (in the domains of
mathematics, medical diagnosis, and determining the structure of complex organic molecules) and suddenly everyone wanted to write an expert system. In the short term, this enthusiasm generated additional funding and research, but it soon became apparent that an expert system in one domain could not necessarily be used as a model for one in another domain. If expert systems showed that real applications had to come to grips with formalizing real knowledge (as opposed to knowledge about toy domains), they also showed that this task was a formidable one. The fourth development was the re-emergence of neural network modelling, of the kind that had been largely set aside by those who espoused the Newell and Simon information processing approach. Theoretical developments together with the availability of larger, faster computers suddenly saw this approach producing important and enticing results.

The 1980s saw the working out of these developments. Although all remain important, all have faced disappointments. It is very hard to make an expert system that replaces an expert, though much easier to write a program that helps one. And it is hard to generalize the lessons learned in one domain of expertise. Cognitive science has not integrated its subdisciplines as closely as was hoped, and neural network modelling has still to show that it can make significant contributions to modelling abilities that call for complex information processing, in particular high-level processes in language understanding and thinking and reasoning.

**KNOWLEDGE REPRESENTATION**

Intelligent behaviour requires information to be stored, either in a short-term store or a long-term store or, more usually, both. One of the primary tasks of AI is therefore to produce an account of how information is represented in an intelligent system.

We know that the human nervous system has many parts, and that those parts probably operate in different ways. Nevertheless, there are many attractions in proposing that all information is stored in the same format. It may not be the form of information storage that differentiates information processing systems, but the nature of the information and the purpose for which it is used. Partly for this reason, many AI researchers have been attracted to the idea that information should be stored using the logical language known as *first order predicate calculus* (FOPC), and extensions of it that incorporate reasoning about time and modality. An additional attraction of this proposal is that, at least in principle, FOPC is computationally tractable: given a FOPC database, other facts implied by that database can be generated automatically. Other systems of representation are either not known to have or known not to have this property.

Unfortunately, although FOPC appears to have desirable properties, in practice it is extremely cumbersome to use. Partly because of the uniformity
of the representation, facts in a large FOPC database can be difficult to find. Similarly, although there is a well-established procedure for drawing inferences from facts in a FOPC database (the resolution method, Robinson, 1965), it very quickly gets bogged down in making all but the simplest inferences. Furthermore, inferences made from a FOPC database cannot be over­ridden by new information. Everyday inferences can – they are said to be non-monotonic. For example, if I know that John is 25 years old and lives in Los Angeles, I infer that he can drive. If I subsequently learn that he suffers from epilepsy, I would probably withdraw my previous conclusion. Since the late 1970s there have been several attempts to construct non-monotonic logics, similar to FOPC but with additional rules of inference that violate monotonicity. There have also been attempts to formalize non-monotonic reasoning in other ways. The idea of a truth maintenance system (TMS) (Doyle, 1979) has been important in many of these. A TMS stores information about the justification for beliefs held, and allows dependency-dependent backtracking, so that when a belief turns out to be false, the reasons why it was held can be accessed directly and reassessed. None of these attempts to handle non-monotonic reasoning has been entirely successful.

Partly as a result of problems with uniform representation systems, such as FOPC, many AI researchers have proposed non-uniform representations, which allow special procedures for manipulating certain types of information. One of the earliest, and best-known, non-uniform representations is semantic networks (Quillian, 1968). Semantic networks give a special place to the information represented in their links and, in particular, they allow efficient processing of taxonomic information. Quillian’s original, and rather simple, networks have been extended and elaborated in various ways, and representation of information in network form has proved a recurrent theme in AI. More complex non-uniform representation schemes that are related to semantic networks include frames and scripts. Scripts represent stereotyped sequences of events, frames have several uses. In one, frames represent particular objects and types of object, and a more recent development is that of object-oriented programming languages. The first widely used object-oriented language was the AI language SMALLTALK. More recently object-oriented versions of the most important AI language, LISP, have appeared, and languages such as C now have object-oriented versions (C++). Indeed, one of the major applications of object-oriented programming is not in AI, but in the development of windows-based interfaces for personal computers and workstations, where windows are treated as objects.

In the framework of semantic networks, the spread of activation through a network is the principal method of extracting information from it. This process has usually been simulated on a serial computer, but it ought to be achieved more efficiently on parallel hardware. Indeed, one of the most important parallel processing computers, the Connection Machine (not to be confused with connectionist neural nets), was inspired by Scott Fahlman’s
(1979) suggestion for implementing semantic networks on special hardware. The idea of distributed processing is also found in neural network models of cognitive processing. Neural networks also allow, though they do not demand, distributed representations of the knowledge embodied in them. In particular, those neural networks that learn to perform tasks, rather than having information encoded into them by the programmer, are likely to develop distributed representations. Such networks show rule-governed behaviour as an emergent property, and the only way to determine exactly what rules such a network is following is to examine the relation between its inputs and its outputs.

There are many things we cannot be sure of, so a further issue in knowledge representation is the encoding and use of uncertain information. Inferences from uncertain information are modelled mathematically using probability theory and, in particular, Bayes’ theorem, which is familiar to psychologists from statistical courses. Complex sets of probabilistic interrelations can be modelled in so-called Bayesian networks. Unfortunately Bayesian inference is neither computationally simple nor always the correct model of real world uncertain inference. The early expert system MYCIN (see below) introduced the simplifying idea of certainty factors associated with each of its diagnostic rules of inference. In recent years attention has focused on a more sophisticated mathematical approach known as Dempster-Shafer theory and there has also been renewed interest in fuzzy set theory, which enjoyed brief popularity in cognitive psychology in the mid-1970s.

VISION

Traditional AI research on vision was concerned, broadly speaking, with recognition of the objects – the prismatic solids – in the BLOCKSWORLD. For computer vision programs, the objects were matt white, uniformly lit (no shadows), and placed against a black background. In fact, the general problem of object recognition in the BLOCKSWORLD was set aside in favour of two of its component problems: finding lines in an image of a BLOCKSWORLD scene, and segmenting the image into sets of regions – each region corresponding to a surface – that belong to the same object. Indeed, this research came to be dominated by attempts to solve the segmentation problem: many programs required line drawings (rather than images) as their inputs.

The most important method of attempting to solve the segmentation problem, originally suggested by Alfonso Guzman (1968), was to use information about the types of vertex in the scene. Guzman’s taxonomy was intuitive, but it was systematized independently by Max Clowes (1971) and David Huffman (1971), who stressed the importance of maintaining different descriptions of the image (in terms of lines, line junctions, and regions) and the scene (in terms of edges, vertices, and surfaces), and of making systematic
inferences about the scene on the basis of the image. The Clowes-Huffman scheme is limited to scenes with no shadows and in which no more than three lines meet at any point. It has three types of line (corresponding to boundaries, inside edges, and outside edges) and four basic types of line junction (Ts, Ys, Ls, and arrows). From these line types and junction types, 16 derived junction types can be constructed, which correspond to possible

![Figure 1](image)

*Figure 1* The 16 derived junction types in the Clowes-Huffman scheme – 4 Ts, 3 Ys, 6 Ls, and 3 arrows. An arrow on a line signifies that it represents an occluding edge (boundary between objects), a plus (+) sign signifies a convex (or outside) edge of a single object, and a minus (−) sign a concave (or inside) edge. The direction of the arrow indicates the side of the line on which the occluding object lies (to the right when facing in the direction of the arrow)
configurations in a BLOCKSWORLD scene (see Figure 1). Identification of the basic junction types in the image, plus the application of the constraint that any line should be of the same type along its whole length, allows most images of permissible scenes to be interpreted.

David Waltz (1975) extended the Clowes-Huffman scheme to scenes with shadows and to images in which more than three lines meet at a point. These apparently simple changes increased the number of permissible derived junction types from 16 to about 2,500. Nevertheless, Waltz's program was more successful than those devised by Clowes and Huffman, since he exploited the need for consistent labelling of neighbouring junctions. An iterative technique known as Waltz filtering or, more generally, as relaxation eliminates possible labellings of junctions, using this consistency constraint. In most cases it rapidly converges on a solution to the segmentation problem for the image it is processing.

Steve Draper (1981) and others have identified a number of problems with the junction-labelling technique and with an alternative to it known as the gradient-space method. Draper invented a technique called sidedness reasoning. Sidedness reasoning is about whether two points or surfaces are on the same side of a third surface. Draper showed that this technique was able to segment all BLOCKSWORLD images but in doing so he virtually put a stop to work on object recognition in the BLOCKSWORLD. The reason was that his technique wore on its sleeve the fact that it was specific to BLOCKSWORLD: it works only when all surfaces are flat. Thus, the idea of solving the problem of object recognition in a miniature domain and scaling up the solution to the real world would not work.

A quite different approach to the problems of vision is found in the work of David Marr (1982) and his associates. Marr's work integrates ideas from AI, psychology, and neurophysiology in what is usually taken to be the paradigmatically successful piece of research in cognitive science. The work is guided by an underlying philosophy about the study of natural information-processing systems. Marr identified three levels at which such systems should be studied. First, a task analysis answers the questions of what the system does and why it does it. This analysis leads to a computational theory of the system — an account of the function (in the mathematical sense) it computes. The second level of analysis is that of representation and algorithm. The third level is that of implementation. In the case of natural information processing systems, this level of analysis requires the study of the neural mechanisms that support the system. Marr is critical of previous AI work on vision, largely because of its focus on the second level of analysis at the expense of the first, to which Marr attached great importance. He is also critical of neurophysiological work, such as that of Hubel and Wiesel (1962), in which the purpose of certain types of cell is inferred from their properties. According to Marr, the purpose of a system (and of its parts) can be determined only by constructing a computational theory.
In his own work, Marr recognized three main stages of visual processing. In the first of these stages, the array of light falling on the retina is transformed into a representation called the *primal sketch*. The primal sketch is a symbolic representation, but it is a representation of the image, not of the scene. It contains information about lines, boundaries, and regions in the image. The construction of the primal sketch takes place very early in the visual system and proceeds on the basis of local interactions between processing units (cells) that represent adjacent parts of the image. Although these interactions reflect what is known about the early visual system, Marr eschewed theories what were motivated *solely* by neurophysiological evidence. Hence, his demand for independent support — from task analysis and psychological evidence — for the algorithm and representation he proposed.

In the second stage of visual processing, the $2\frac{1}{2}$D sketch is derived from the primal sketch. This sketch is a very short-term memory store into which a set of processes writes information about the surfaces (in the scene) represented in the image, their orientation, and their approximate distance from the viewer: the third dimension is not properly represented, hence $2\frac{1}{2}$D sketch. The most important of these processes are stereopsis, structure from motion, and shape from shading.

Since objects have not yet been recognized, surfaces cannot be identified by reference to information about the objects of which they are part. This aspect of the construction of the $2\frac{1}{2}$D sketch reflects Marr’s preference for *bottom-up* (data-driven) theories of visual processing. The only world knowledge that such theories can claim the visual system uses is a set of general principles, such as what very few points in an image correspond to abrupt changes in the surface represented. Specific information about the scene being viewed is not yet available.

In the final stage of visual processing, a *3D model description* is constructed from the $2\frac{1}{2}$D sketch. This representation contains information about the identity and three-dimensional structure of the objects in the scene. Marr’s account of this final stage is highly speculative, and less closely linked with the psychological and neurophysiological evidence. Marr’s basic idea is that objects can be represented, in a *catalogue* stored in long-term memory, as jointed *generalized cylinders* (cylinders whose cross-section changes along their length). The principal axes of these cylinders make up stick figures of the objects represented. He showed that, subject to certain constraints, generalized cylinder representations could be derived from the $2\frac{1}{2}$D sketch, and then compared with entries in the catalogue, with any necessary rotation and bending at the joints. In practice this matching is difficult, and Marr suggested a process of gradual refinement in the match between the image and the stored representations in the catalogue. This kind of process can be (relatively) time-consuming, and was rejected by Marr in his analyses of the lower levels of visual processing.
Marr's work incorporates, in addition to traditional AI-style programming, much straightforward mathematics. Subsequent work on vision, both theoretical and applied, has become increasingly mathematical and, hence, increasingly inaccessible to psychologists. On the theoretical side, many of the problems of visual analysis have been identified as special cases of what are known as *ill-posed* problems. They are ill posed because, as they stand, they do not have a unique solution. They can be analysed by a technique known as *regularization*, which requires the addition to the problem of the kind of general constraints identified by Marr. On the applied side, specialized hardware in the form of very large-scale integration (VLSI) chips has allowed, for example, stereo algorithms to be used in real-world applications.

**THINKING, REASONING, PROBLEM SOLVING**

Historically, problem solving was one of the earliest topics of AI research. Furthermore, it has often been argued that it is the central topic, since AI techniques in other domains can be seen as special cases of searching through a "space" of possibilities for a solution to a problem. For example, parsing a sentence can be seen as a search through the (infinite) set of possible syntactic structures defined by the grammar of a language.

Occasionally it is possible to examine all possible solutions to a problem to find the right one. However, for most interesting problems there are too many possibilities to make this approach viable. Usually there are several steps in the solution to a problem, so the number of possible moves multiplies up at each step, producing what is called a *combinatorial explosion* in the number of potential solutions. A *control strategy* for searching through the space of possible solutions is, therefore, required.

Traditionally, there are two ways of representing problems so that a search can be made for their solution. In a *state-space representation*, problems are represented in terms of states of the relevant part of the world, and actions (usually referred to as *operators*) that transform one state into another. In this representation, a single path through the tree of possibilities (= a sequence of operators) represents the solution to the problem. In a *problem-reduction* representation a large problem is broken up into a number of sub-problems, all of which must be solved if the main problem is to be solved. State-space representations are easier to construct. Sensible reductions of problems can be hard to find, but they are very useful when they have been found. In serious AI work on problem solving the two types of representation are combined into AND/OR trees. AND branchings represent problem reductions, where all the sub-goals have to be fulfilled. OR branchings represent alternative possibilities in a state space, only one of which has to be fulfilled.

Various general control strategies for searching problem spaces have been proposed. The most fundamental distinction is between *breadth-first* and
**ARTIFICIAL INTELLIGENCE**

*depth-first* search of trees. In breadth-first search all possible one-operator solutions are checked, then all possible two-operator solutions, and so on. In depth-first search one possible solution is followed up until it succeeds or fails, or until a pre-set depth limit is reached, since a branch in an AND/OR tree may never terminate. Simple depth-first and breadth-first search are used only in desperation. Usually some method is introduced for following up the most promising possibilities. Methods for deciding which possibility is the most promising are inevitably heuristic. The most sophisticated method of making the choice is the AO* algorithm. However, the algorithm itself does not provide the means of measuring which next move is the best. Furthermore, there is no general method for assigning values to moves. A new one must be devised for each domain in which the algorithm is used.

Such methods can, nevertheless, be applied to solving puzzle-book problems and in game-playing computers (e.g., for chess). In chess-playing programs the problem that the computer is trying to solve is not how to win the game, but what move to make next. Successful programs run on very fast super-computers, so that they can examine vast numbers of possible moves. However, they limit the distance ahead (in terms of moves) that they look. Since they typically cannot see ahead to a winning position, they have to evaluate the positions that they can reach in other ways, and then aim to reach the best position that a rational opponent will let them. The play of such programs differs in several ways from that of human chess players. The standard of the best of them, however, is usually reckoned to be in the grandmaster category.

Even if all AI researchers had access to the kind of super-computers that chess programmers use, they would not necessarily want to use the same kind of brute force problem solving methods, particularly if they were interested in modelling human problem solving abilities. Newell, Shaw, and Simon (1957) first introduced the idea of heuristic (rule-of-thumb) problem solving techniques in their Logic Theory Machine, that proved theorems of logic. An alternative way of speeding up problem solving is to use domain-specific techniques, that may be heuristic, but which need not be. An early example of an AI program that used a domain-specific technique was Gelernter’s (1963) Geometry Machine, which constructed the equivalent of geometrical diagrams. It is thought that most human mathematicians, except when they are working in completely new areas of mathematics, use domain-specific techniques. More generally, domain-specific techniques are thought to be widely used in all types of problem solving.

**LANGUAGE**

There is a long history of computational research on all aspects of language processing. Research on speech, both automatic speech recognition and speech synthesis, has been strongly influenced by work on signal processing.
carried out by electronic engineers. More recently, with the advent of larger and more powerful computers, the field of speech and language technology has emerged, which is primarily directed to producing tools for processing large corpora of linguistic data held on computers. Some of the techniques developed may be of interest to AI researchers; others are used to derive statistical information that is of primary interest to, say, lexicographers.

Work on language processing is divided into three parts, concerned respectively with recognizing or selecting words, computing or generating sentence structure, and processing meaning at the level of discourse. Until the 1970s AI research on language processing often produced working systems that understood a substantial portion of a language such as English. Winograd's (1972) SHRDLU, a program that talks about moving blocks around the BLOCKSWORLD, represents the apotheosis of this work. However, it has since become obvious that the component parts of language processing are each so complex that they must be studied separately, if real progress is to be made.

Recent work on word identification has been largely dominated by neural network modelling, in particular the TRACE model of auditory word identification (McClelland & Elman, 1986) and Seidenberg and McClelland's (1989) model of visual word identification. The TRACE model is "hand-coded". It does not use distributed representations, and hence its mode of operation is easy to discern. It has interacting banks of detectors at three levels: for the auditory features of sounds, for phonemes (sounds that correspond roughly to letters), and for words. The Seidenberg and McClelland model, on the other hand, is a model that learns. One of its most interesting features is its eschewal of lexical representations: all its knowledge is encoded in links between orthographic and phonological features.

Investigations of the computation of sentence structure (parsing) have taken two rather different directions. On the one hand, computational linguists worry about problems such as the linguistic niceties of describing sentence structure and the computational properties of the procedures that derive the structure for a particular sentence, given a description of how sentences in its language can be structured (a grammar). One of the most important developments in computational models of parsing is the introduction of unification-based approaches (e.g., Kay, 1985). Unification is a technique that is widely used in other branches of AI, in particular theorem proving. Unification-based parsers, like some other parsers, such as chart parsers, have the additional advantage of clearly separating information about how sentences can be structured (the grammar) from information about how sentence structure is computed (the parsing algorithm). In contrast with researchers whose primary interest is in the computational properties of parsing systems, those who attempt to model the way that people derive sentence structure have to take account of well-established empirical findings on, in particular, what happens when people encounter a syntactic ambiguity. It is not yet clear how these two approaches to parsing can be integrated.
Understanding and generating discourse still remain formidable tasks. AI research has often been hampered by a restricted or ad-hoc approach to word meanings. One hope is that linguistically more sophisticated approaches to word meaning, such as Jackendoff’s (1990) conceptual semantics, will be taken up by AI researchers. At the level of sentence meaning, AI researchers, at last, agree about the importance of compositional semantics of a broadly Montagovian kind (Dowty, Wall, & Peters, 1981). However, the major problems in describing discourse level processing, which have been known for many years, still resist satisfactory analysis. Some of the most important are figurative and indirect uses of language, coherence, ellipsis, and the role of the other participants’ beliefs.

LEARNING

For historical reasons, learning has been a comparatively neglected topic in AI. The information processing approach to understanding intelligent behaviour was seen as a radical alternative to the behaviourism that had dominated psychology, and which placed a strong emphasis on learning. Furthermore, traditional AI aimed to study intelligence at an abstract level, independent of both its genesis (learned or programmed) and its underlying mechanism (carbon or silicon). The study of learning has come back into its own with the increasing importance of connectionist modelling. Nevertheless, a number of important studies of learning have been carried out in the symbolic framework, and the diversity of the learning mechanisms that they investigate contrasts sharply with the behaviourist approach.

Learning by being told often involves little more than adding a fact to a database. However, more abstract pieces of information, such as advice on the best strategy for winning a game, may need to be operationalized.

A more complex kind of learning is learning from mistakes. Gerald Sussman’s (1975) program HACKER writes its own mini-programs for solving problems of stacking and unstacking blocks in BLOCKSWORLD. However, it can learn only when it can almost solve a problem, and its performance is crucially dependent on its having a “teacher” who presents it with a suitably graded set of problems. Patrick Winston’s (1975) program that learns concepts for configurations of blocks (such as arches) in BLOCKSWORLD, similarly learns from almost correct information. When told that something is not quite an arch, it can use that information to deduce what distinguishes arches from non-arches.

As well as recognizing the importance of being almost correct, Winston also emphasized that an important aspect of learning is what is sometimes called induction – going beyond the information embodied in the examples presented to the program to form general concepts (in his case) or rules. Positive instances suggest generalizations of the concept or rule, negative instances suggest specializations (or restrictions). Research subsequent to
Winston’s, particularly that of Ryszard Michalski (e.g., 1983) has systematized the study of induction, and shown that it can be regarded as a special case of search, with the search space being the set of possible generalizations statable in a particular language. Michalski’s approach is more powerful than Winston’s, but less closely related to human learning. It can also be used for the related task of discrimination learning. Its disadvantage is that it works straightforwardly only if the generalizations are formulated using exactly the same predicates that are used to describe the instances.

Winston’s program can learn more complex concepts (such as arch) only because it knows simpler concepts (pillar, lintel). This aspect of the program relates, very crudely, to the question of how much of what we know about language is learned, and how much is innate. In the case of concepts, it has been argued (e.g., by Fodor, 1981) that all concepts must be innate. More generally, it is widely, though not universally, believed that many general principles governing the form of possible languages are innate, and that the availability of these principles to the language learning mechanism explains how it is able to achieve what appears, on mathematical analysis, to be a difficult or impossible task.

Another famous example of learning by generalization is Arthur Samuel’s (1963) checkers (draughts) program. This program develops a general method for evaluating board positions by comparing computed evaluations with the way the game actually turns out, and revising, if necessary, the method of evaluation.

A more ambitious, and more controversial, attempt to study a different kind of learning – learning by exploration – is found in Doug Lenat’s (1982) AM (Automated Mathematician) and EURISKO programs. AM starts with a collection of set-theoretic concepts and ways of combining them, and creates further mathematical concepts from them (e.g., positive whole number, prime number, the fundamental theorem of arithmetic – that every number can be expressed as a product of prime factors).

None of the programs described so far provides a convincing model of human learning. People can learn things very quickly, though they often make mistakes in doing so. This very quick learning depends on particular ways of using background knowledge. Two lines of research that attempt to model this kind of learning investigate analogy-based learning and explanation-based learning. The importance of analogy in learning and problem-solving has long been recognized in cognitive psychology. None the less the underlying processes are difficult to model computationally, not least because the domain from which an analogy is drawn need not be specified in advance. In explanation-based learning (see e.g., de Jong, 1988) a single event or episode is explained on the basis of a theory about the relevant aspects of the world. That explanation is then generalized so that it will be useful in other situations.

Traditional AI work on learning has embodied a variety of ideas. An
alternative tradition, running from the British Empiricist philosophers of the seventeenth and eighteenth centuries to the behaviourists and neo-behaviourists of the twentieth century, has seen all learning as the formation and strengthening of associations between ideas. In a modified form, this notion also underlies recent connectionist accounts of learning. Connectionists machines are collections of simple processing units, with levels of

![A simple connectionist network showing the three types of unit — input, hidden, and output — and the connections between them](image)

**Figure 2** A simple connectionist network showing the three types of unit — input, hidden, and output — and the connections between them
activation that can be passed from one unit to another. A typical machine has three layers of units: input units, hidden units and output units (see Figure 2). Such machines can learn in several ways, but the most popular is known as a *back propagation*. It is a supervised learning method in which a stimulus is encoded at the input units and produces an output at the output units. The supervisor tells the machine what the output should have been, and the difference between the actual and expected outputs is propagated back through the network of units, and used, in a precisely specified way, to adjust the (associative) strengths of the connections between them. Adjustments are small, because the machine must not produce the correct response to the last input at the expense of responding grossly incorrectly to other inputs. Learning is slow, sometimes very slow, but a stable set of associative strengths is usually reached.

Another biological metaphor that has inspired AI work on learning is *evolution*. Genetic algorithms (e.g. Goldberg, 1989) use complex rules to perform tasks. The parts of these rules can be recombined by processes that are analogous to the genetic operations that take place in the germ cells during sexual reproduction. The resulting rules are then allowed to perform their task for some time, and their performance is assessed. Those that do best re-enter the “reproductive” process.

**APPLICATIONS**

Intelligent machines should be of more than academic interest. However, most of the machines that we interact with in everyday life, for example automatic bank tellers, are not intelligent. More intelligent machines — often referred to as expert systems — do have applications. However, despite the hopes of the early 1980s, it now appears that expert systems will typically be used to assist experts, rather than to replace them. Perhaps the most important area of application for intelligent programs is in medical diagnosis, though there are obviously ethical problems in this domain. One area in which computers play a crucial role is in modern scanning techniques (CAT, PET, NMR, etc.). The basic use of computers in scanning is to generate appropriate images. Intelligent programs might also help to produce diagnoses from images.

One of the earliest, and best known, medical diagnosis systems is MYCIN (Shortliffe, 1976), which diagnoses serious bacterial infections so that life-saving antibiotic drugs can be administered before a culture has been developed. The development of such a system requires the gleaning of information about the diseases in question and their symptoms. Some of this information is elicited from experts, sometimes with difficulty, as the experts cannot necessarily verbalize their knowledge. TEIRESIAS (Davis, 1982) is a program that attempts to automate this knowledge transfer, and also to use the knowledge already in MYCIN to generate user-friendly explanations of
ARTIFICIAL INTELLIGENCE

its diagnoses. Other diagnostic information comes from statistical records. In an expert system all the information is usually represented in a uniform way, so that new information can readily be added. The rules for making inferences are stored separately, and an attempt is made to keep the inferential processes simple. One of the major aspects of inference in expert systems is combining uncertain bits of information to produce a best guess, for example at a diagnosis. This combination is sometimes achieved using standard statistical (Bayesian) techniques and sometimes using domain specific rules, as in MYCIN (see above).

MYCIN also formed the basis of the first expert system shell, E-MYCIN, which is MYCIN stripped of its domain-specific knowledge. Expert system shells were the first of several attempts to make the creation of new expert systems easy. Success has been partial. E-MYCIN, for example, is most successful in other medical diagnosis systems, such as PUFF, which diagnoses pulmonary diseases.

Another well-known expert system is DENDRAL (Lindsay, Buchanan, Feigenbaum, & Lederberg, 1980), which works out the molecular structure of large organic molecules from their mass spectrograms. DENDRAL has been in regular use by research chemists for some time. An additional program, meta-DENDRAL, attempts to formulate new rules using the induction techniques described above.

A second area in which AI has sought to find application is in computer-assisted learning (CAL). With the expansion of higher education in the UK, CAL is likely to become increasingly important, though it is as yet unclear what the contribution of AI techniques will be. The current focus of attention is on multimedia, and in particular hypermedia learning tools, which provide facilities for exploring large databases in various ways, but which rely on much of the intelligence resting in the instructions and with the student.

The intelligent tutoring systems of AI, on the other hand, try to be intelligent themselves. Such systems have three main components: a knowledge base which could, in principle, incorporate multimedia options, a model of the student, and a set of teaching strategies. The knowledge base is used to impart information directly to students, but it is also used to generate explanations of why students’ answers to questions are wrong. This process, in turn, makes use of the model of the student to decide what kinds of misconceptions students will have. Such indirect methods of teaching meet with some success, but they prove comparatively difficult to implement in a tutoring system.

PHILOSOPHICAL ISSUES

AI research, more than that in other sciences, has been surrounded by philosophical controversy. Two related issues have provided the major focus
of debate. The first is whether machines can think, and the second is what role they should be allowed to play in our lives.

The question of whether machines can think, although one that excites the popular imagination, is not necessarily a clear one. One crucial aspect of it, however, is whether there is a difference between computer programs that model phenomena such as the weather, which simulate processes in the world, but do not reproduce them, and AI programs. In other words: is a computer running such a program really intelligent, or is it just simulating intelligent behaviour? On one view, most programs lack real intelligence because they do not interact with the world. The symbols that they manipulate have meaning only because of the way they are interpreted by their programmers. On this view a robot that based its interactions with the world on its internal computations could be intelligent. An opposing view is that real intelligence can be manifest only in biological systems (Searle, 1980). To support this thesis Searle put forward his famous Chinese room argument. If he sat in a room manipulating symbols according to the rules embodied in a computer program, he might, from the outside, be described as reading and answering questions in Chinese. He would not, however, understand Chinese. So, understanding Chinese is not just running a program. However, Searle’s view of what else it is, basically being a biological intelligence, appears to have no foundation, and has been dubbed protoplasm chauvinism (Torrance, 1986).

If machines, or at least robots, can be intelligent, we might at some time in the future have moral responsibilities towards them, or we might be in danger of being dominated by them. To some extent the moral issues raised by such considerations are just those that arise in the application of any science. The difference is that we might be faced not simply with a substance or technique that might be misused, but with something that is itself an “alien” intelligence. However, it is difficult to pinpoint, as Weizenbaum (1976) has tried to do, the sense in which intelligent computers pose a special threat.

**ARTIFICIAL INTELLIGENCE, COGNITIVE PSYCHOLOGY, AND THE FUTURE**

Since the mid-1970s there has been an enormous growth in AI research. It is no longer possible, as it once was, for an AI researcher, let alone a psychologist, to keep up with developments in all of its subfields. Furthermore, much of AI has become very technical: much more so than cognitive psychology. Nevertheless, the best science often is technical; if cognitive psychologists are not to risk being usurped, they should keep at least one eye on developments in AI.
REFERENCES


