

11/10/02b

newspaper accident reports, had answered my challenge that a computer would count as intelligent only if it could summarize a short story.³ But Schank's newspaper program cannot provide a clue concerning judgments of what to include in a story summary because it works only where relevance and significance have been predetermined, and thereby avoids dealing with the world built up in a story in terms of which judgments of relevance and importance are made.

Nothing could ever call into question Schank's basic assumption that all human practice and know-how is represented in the mind as a system of beliefs constructed from context-free primitive actions and facts; but there are signs of trouble. Schank does admit that an individual's "belief system" cannot be fully elicited from him—though he never doubts that it exists and that it could in principle be represented in his formalism. He is therefore led to the desperate idea of a program which could learn about everything from restaurants to life themes the way people do. In one of his papers he concludes:

We hope to be able to build a program that can learn, as a child does, how to do what we have described in this paper, instead of being spoon-fed the tremendous information necessary.

(1972, pp. 553–554)

In any case, Schank's appeal to learning is at best another evasion. Developmental psychology has shown that children's learning does not consist merely in acquiring more and more information about specific routine situations by adding new primitives and combining old ones, as Schank's view would lead one to expect. Rather, learning of specific details takes place on a background of shared practices which seem to be picked up in everyday interactions not as facts and beliefs but as bodily skills for coping with the world. Any learning presupposes this background of implicit know-how which gives significance to details. Since Schank admits that he cannot see how this background can be made explicit so as to be given to a computer, and since the background is presupposed for the kind of script learning Schank has in mind, it seems that his project of using preanalyzed primitives to capture common sense understanding is doomed.

2.3 KRL: a knowledge-representation language

Winograd and Bobrow propose a more plausible, even if in the last analysis perhaps no more promising, approach that would use the new theoretical power of frames or stereotypes to dispense with the need to

preanalyze everyday situations in terms of a set of primitive features whose *relevance is independent of context*. This approach starts with the recognition that in everyday communication: "Meaning' is multidimensional, formalizable only in terms of the entire complex of goals and knowledge [of the world] being applied by both the producer and understander." (Winograd 1976b, p. 262) This knowledge, of course, is assumed to be "a body of specific beliefs (expressed as symbol structures ...) making up the person's 'model of the world'" (p. 268). Given these assumptions, Winograd and his coworkers are developing a new knowledge-representation language (KRL), which they hope will enable programmers to capture these beliefs in symbolic descriptions of multidimensional prototypical objects whose *relevant aspects are a function of their context*.

Prototypes would be structured so that any sort of description from proper names to procedures for recognizing an example could be used to fill in any one of the nodes or slots that are attached to a prototype. This allows representations to be defined in terms of each other, and results in what the authors call "a *holistic* as opposed to *reductionistic* view of representation" (Bobrow and Winograd 1977, p. 7). For example, since any description could be part of any other, chairs could be described as having aspects such as seats and backs, and seats and backs in turn could be described in terms of their function in chairs. Furthermore, each prototypical object or situation could be described from many different perspectives. Thus nothing need be defined in terms of its necessary and sufficient features in the way Winston and traditional philosophers have proposed, but rather, following Rosch's research on prototypes, objects would be classified as more or less resembling certain prototypical descriptions.

Winograd illustrates this idea by using the traditional philosophers' favorite example:

The word 'bachelor' has been used in many discussions of semantics, since (save for obscure meanings involving aquatic mammals and medieval chivalry) it seems to have a formally tractable meaning which can be paraphrased "an adult human male who has never been married" ... In the realistic use of the word, there are many problems which are not as simply stated and formalized. Consider the following exchange.

Host: I'm having a big party next weekend. Do you know any nice bachelors I could invite?

Yes, I know this fellow X.

The problem is to decide, given the facts below, for which values of X the response would be a reasonable answer, in light of the normal meaning of the word "bachelor". A simple test is to ask for which ones the host might fairly complain "You lied. You said X was a bachelor".

- A: Arthur has been living happily with Alice for the last five years. They have a two year old daughter and have never officially married.
- B: Bruce was going to be drafted, so he arranged with his friend Barbara to have a justice of the peace marry them so he would be exempt. They have never lived together. He dates a number of women, and plans to have the marriage annulled as soon as he finds someone he wants to marry.
- C: Charlie is 17 years old. He lives at home with his parents and is in high school.
- D: David is 17 years old. He left home at 13, started a small business, and is now a successful young entrepreneur leading a playboy's life style in his penthouse apartment.
- E: Eli and Edgar are homosexual lovers who have been living together for many years.
- F: Faisal is allowed by the law of his native Abu Dhabi to have three wives. He currently has two and is interested in meeting another potential fiancée.
- G: Father Gregory is the bishop of the Catholic cathedral at Groton upon Thames.

[This] cast of characters could be extended indefinitely, and in each case there are problems in deciding whether the word 'bachelor' could appropriately be applied. In normal use, a word does not convey a clearly definable combination of primitive propositions, but evokes an *exemplar* which possesses a number of properties. This exemplar is not a specific individual in the experience of the language user, but is more abstract, representing a conflation of typical properties. A prototypical bachelor can be described as:

1. a person
2. a male
3. an adult
4. not currently officially married
5. not in a marriage-like living situation

6. potentially marriageable
7. leading a bachelor-like life style
8. not having been married previously
9. having an intention, at least temporarily, not to marry
10. ...

Each of the men described above fits some but not all of these characterizations. Except for narrow legalistic contexts, there is no *significant sense* in which a subset of the characteristics can be singled out as the "central meaning" of the word. In fact, among native English speakers there is little agreement about whether someone who has been previously married can properly be called a "bachelor" and fairly good agreement that it should not apply to someone who is not potentially marriageable (for instance, has taken a vow of celibacy).

Not only is this list [of properties] open-ended, but the individual terms are themselves not definable in terms of primitive notions. In reducing the meaning of 'bachelor' to a formula involving 'adult' or 'potentially marriageable', one is led into describing these in terms of exemplars as well. 'Adult' cannot be defined in terms of years of age for any but technical legal purposes and in fact even in this restricted sense, it is defined differently for different aspects of the law. Phrases such as 'marriage-like living situation' and 'bachelor-like life-style' reflect directly in the syntactic form the intention to convey stereotyped exemplars rather than formal definitions. (1976b, 276-278)

Obviously, if KRL succeeds in enabling AI researchers to use such prototypes to write flexible programs, such a language will be a major breakthrough and will avoid the *ad hoc* character of the "solutions" typical of micro-world programs. Indeed, the future of AI depends on some such work as that begun with the development of KRL. But there are problems with this approach. Winograd's analysis has the important consequence that in comparing two prototypes, what counts as a match and thus what count as the relevant aspects which justify the match will be a result of the program's understanding of the current context.

The result of a matching process is not a simple true/false answer. It can be stated in its most general form as: "Given the set of alternatives which I am currently considering ... and looking in order at those stored structures which are most accessible in the *current context*, here is the best match, here is the degree to which it seems

to hold, and here are the specific detailed places where match was not found ...”

The selection of the order in which substructures of the description will be compared is a function of their current accessibility, which depends both on the form in which they are stored and the *current context*. (1976b, p. 281–282; emphasis added)

This raises four increasingly grave difficulties. *First*, for there to be “a class of cognitive ‘matching’ processes which operate on the descriptions (symbol structures) available for two entities, looking for correspondences and differences” (p. 280), there must be a finite set of prototypes to be matched. To take Winograd’s example:

A single object or event can be described with respect to several prototypes, with further specifications from the perspective of each. The fact that last week *Rusty flew to San Francisco* would be expressed by describing the event as a typical instance of *Travel* with the mode specified as *Airplane*, destination *San Francisco*, and so on. It might also be described as a *Visit* with the actor being *Rusty*, the friends a particular group of people, the interaction warm, and so on. (Bobrow and Winograd 1977, p. 8)

But “*and so on*” covers what might, without predigestion for a specific purpose, be a hopeless proliferation. The same flight might also be a test flight, a check of crew performance, a stopover, a mistake, a golden opportunity, not to mention a visit to brother, sister, thesis adviser, guru, *and so on, and so on, and so on*. Before the program can function at all, the total set of possible alternatives must be pre-selected by the programmer.

Second, the matching makes sense only *after* the current candidates for comparison have been found. In chess, for example, positions can be compared only after the chess master calls to mind past positions that the current board positions might plausibly resemble. And (as in the chess case) the discovery of the relevant candidates which make the matching of aspects possible requires experience and intuitive association.

The only way a KRL-based program (which must use symbolic descriptions) could proceed, in chess or anywhere else, would be to guess some frame on the basis of what was already “understood” by the program, and then see if that frame’s features could be matched to some current description. If not, the program would have to backtrack and try another prototype until it found one into whose slots or default terminals the incoming data could be fitted. This seems an

altogether implausible and inefficient model of how we perform, and only rarely occurs in our conscious life. Of course, cognitive scientists could answer the above objection by maintaining, in spite of the implausibility, that we try out the various prototypes very quickly and are simply not aware of the frantic shuffling of hypotheses going on in our unconscious. But, in fact, most would still agree with Winograd’s (1974) assessment that the frame selection problem remains unsolved.

The problem of choosing the frames to try is another very open area. There is a selection problem, since we cannot take all of our possible frames for different kinds of events and match them against what is going on. (p. 80)

There is, moreover, a *third* and more basic question which may pose an in-principle problem for any formal holistic account in which the significance of any fact, indeed what counts as a fact, always depends on the context. Bobrow and Winograd stress the critical importance of context.

The results of human reasoning are *context dependent*, the structure of memory includes not only the long-term storage organization (What do I know?) but also a current context (What is in focus at the moment?). We believe that this is an important feature of human thought, not an inconvenient limitation. (1977, p. 32)

Winograd further notes that “the problem is to find a formal way of talking about ... current attention focus and goals” (1976b, p. 283). Yet he gives no formal account of how a computer program written in KRL could determine the current context.

Winograd’s work does contain suggestive claims, such as his remark that “the procedural approach formalizes notions like ‘current context’ ... and ‘attention focus’ in terms of the processes by which cognitive state changes as a person comprehends or produces utterances” (pp. 287–288). There are also occasional parenthetical references to “current goals, focus of attention, set of words recently heard, and so on” (p. 282). But reference to recent words has proven useless as a way of determining what the current context is, and reference to current goals and focus of attention is vague and perhaps even question-begging. If a human being’s current goal is, say, to find a chair to sit on, his current focus might be on recognizing whether he is in a living room or a warehouse. He will also have short-range goals like finding the walls, longer-range goals like finding the light switch, middle-range goals like wanting to write or rest; and what counts as satisfying these

goals will in turn depend on his ultimate goals and interpretation of himself as, say, a writer, or merely as easily exhausted and deserving comfort. So Winograd's appeal to "current goals and focus" covers too much to be useful in determining what specific situation the program is in.

To be consistent, Winograd would have to treat each type of situation the computer could be in as an object with *its* prototypical description; then in recognizing a specific situation, the situation or context in which *that* situation was encountered would determine which foci, goals, and the like, were relevant. But where would such a regress stop? Human beings, of course, don't have this problem. They are, as Heidegger puts it, *always already* in a situation, which they constantly revise. If we look at it genetically, this is no mystery: We can see that human beings are gradually trained into their cultural situation on the basis of their embodied precultural situation, in a way no programmer using KRL is trying to capture. But for this very reason a program in KRL is *not* always-already-in-a-situation. Even if it represents all human knowledge in its stereotypes, including all possible types of human situations, it represents them from the outside, like a Martian or a god. It isn't situated *in* any one of them, and it may be impossible to program it to behave as if it were.

This leads to my *fourth* and final question. Is the know-how that enables human beings constantly to sense what specific situation they are in the sort of know-how that can be represented as a kind of knowledge in *any* knowledge-representation language no matter how ingenious and complex? It seems that our sense of our situation is determined by our changing moods, by our current concerns and projects, by our long-range self-interpretation and probably also by our sensory-motor skills for coping with objects and people—skills we develop by practice without ever having to represent to ourselves our body as an object, our culture as a set of beliefs, or our propensities as situation-action rules. All these uniquely human capacities provide a "richness" or a "thickness" to our way of being-in-the-world and thus seem to play an essential role in situatedness, which in turn underlies all intelligent behavior.

There is no reason to suppose that moods, mattering, and embodied skills can be captured in any formal web of belief; and except for Kenneth Colby, whose view is not accepted by the rest of the AI community, no current work assumes that they can. Rather, all AI workers and cognitive psychologists are committed, more or less lucidly, to the

view that such noncognitive aspects of the mind can simply be ignored. This belief that a significant part of what counts as intelligent behavior can be captured in purely cognitive structures defines cognitive science and is a version of what I call the *psychological assumption* (1972/92, chapter 4). Winograd makes it explicit.

AI is the general study of those aspects of cognition which are common to all physical symbol systems, including humans and computers. (see Schank et al. 1977, p. 1008)

But this definition merely delimits the field; it in no way shows there is anything to study, let alone guarantees the project's success.

Seen in this light, Winograd's grounds for optimism contradict his own basic assumptions. On the one hand, he sees that a lot of what goes on in human minds cannot be programmed, so he only hopes to program a significant part.

[C]ognitive science ... does not rest on an assumption that the analysis of mind as a physical symbol system provides a *complete* understanding of human thought ... For the paradigm to be of value, it is only necessary that there be *some significant aspects* of thought and language which can be profitably understood through analogy with other symbol systems we know how to construct. (1976b, p. 264)

On the other hand, he sees that human intelligence is "holistic" and that meaning depends on "the entire complex of goals and knowledge". What our discussion suggests is that all aspects of human thought, including nonformal aspects like moods, sensory-motor skills, and long-range self-interpretations, are so interrelated that one cannot substitute an abstractable web of explicit beliefs for the whole cloth of our concrete everyday practices.

What lends plausibility to the cognitivist position is the conviction that such a web of beliefs must finally fold back on itself and be complete, since we can know only a finite number of facts and procedures describable in a finite number of sentences. But since facts are discriminated, and language is used, only in a context, the argument that the web of belief must in principle be completely formalizable does not show that such a belief system can account for intelligent behavior. This would be true only if the context could also be captured in the web of facts and procedures. But if the context is determined by moods, concerns, and skills, then the fact that our beliefs can in principle be completely represented does not show that representations are

sufficient to account for cognition. Indeed, if nonrepresentable capacities play an essential role in situatedness, and the situation is presupposed by all intelligent behavior, then the "aspects of cognition which are common to all physical symbol systems" will not be able to account for any cognitive *performance* at all.

In the end, the very idea of a holistic information-processing model in which the relevance of the facts depends on the context may involve a contradiction. To recognize any context one must have already selected from the indefinite number of possibly discriminable features the possibly relevant ones; but such a selection can be made only after the context has already been recognized as similar to an already analyzed one. The holist thus faces a vicious circle: relevance presupposes similarity and similarity presupposes relevance. The only way to avoid this loop is to be always-already-in-a-situation without representing it, so that the problem of the priority of context and features does not arise, or else to return to the reductionist project of preanalyzing all situations in terms of a fixed set of possibly relevant primitives—a project which has its own practical problems, as our analysis of Schank's work has shown, and, as we shall see in the conclusion, may have its own internal contradiction as well.

Whether this is, indeed, an in-principle obstacle to Winograd's approach, only further research will tell. Winograd himself is admirably cautious in his claims.

If the procedural approach is successful, it will eventually be possible to describe the mechanisms at such a level of detail that there will be a verifiable fit with many aspects of detailed human performance ... but we are nowhere near having explanations which cover language processing as a whole, including meaning.

(1976b, p. 297)

If problems do arise because of the necessity in any formalism of isolating beliefs from the rest of human activity, Winograd will no doubt have the courage to analyze and profit from the discovery. In the meantime everyone interested in the philosophical project of cognitive science will be watching to see if Winograd and company can produce a moodless, disembodied, concernless, already-adult surrogate for our slowly-acquired situated understanding.

3 Conclusion

Given the fundamental supposition of the information-processing approach that all that is relevant to intelligent behavior can be formalized in a structured description, all problems must appear to be merely problems of complexity. Bobrow and Winograd put this final faith very clearly at the end of their description of KRL.

The system is complex, and will continue to get more so in the near future ... [W]e do not expect that it will ever be reduced to a very small set of mechanisms. Human thought, we believe, is the product of the interaction of a fairly large set of interdependent processes. Any representation language which is to be used in modeling thought or achieving "intelligent" performance will have to have an extensive and varied repertoire of mechanisms.

(Bobrow and Winograd 1977, p. 43)

Underlying this mechanistic assumption is an even deeper assumption which has gradually become clear during the past ten years of research. During this period, AI researchers have consistently run up against the problem of representing everyday context. Work during the first five years (1967-1972) demonstrated the futility of trying to evade the importance of everyday context by creating artificial gamelike contexts preanalyzed in terms of a list of fixed-relevance features. More recent work has thus been forced to deal directly with the background of common-sense know-how which guides our changing sense of what counts as the relevant facts. Faced with this necessity, researchers have implicitly tried to treat the broadest context or background as an object with its own set of preselected descriptive features. This assumption, that the background can be treated as just another object to be represented in the same sort of structured description in which everyday objects are represented, is essential to our whole philosophical tradition. Following Heidegger, who is the first to have identified and criticized this assumption, I will call it the *metaphysical assumption*.

The obvious question to ask in conclusion is: Is there any evidence, besides the persistent difficulties and history of unfulfilled promises in AI, for believing that the metaphysical assumption is unjustified? It may be that no argument can be given against it, since facts put forth to show that the background of practices is unrepresentable are in that very act shown to be the sort of facts which *can* be represented. Still, I will attempt to lay out the argument which underlies my antiformalist, and therefore, antimechanist convictions.

My thesis, which owes a lot to Wittgenstein (1953), is that whenever human behavior is analyzed in terms of rules, these rules must always contain a *ceteris paribus* condition, that is, they apply "everything else being equal"; and what "everything else" and "equal" mean in any specific situation can never be fully spelled out without a regress. Moreover, the *ceteris paribus* condition is not merely an annoyance which shows that the analysis is not yet complete and might be what Husserl called an "infinite task". Rather the *ceteris paribus* condition points to a background of practices which are the condition of the possibility of all rule-like activity. In explaining our actions we must always sooner or later fall back on our everyday practices and simply say "this is what we do" or "that's what it is to be a human being". Thus in the last analysis all intelligibility and all intelligent behavior must be traced back to our sense of what we *are*, which is, according to this argument, necessarily, on pain of regress, something we can never explicitly *know*.

Still, to this dilemma the AI researchers might plausibly respond: "Whatever background of shared interests, feelings, and practices is necessary for understanding specific situations, that knowledge *must* somehow be represented in the human beings who have that understanding. And how else could such knowledge be represented but in some explicit data structure?" Indeed, the kind of computer programming accepted by all workers in AI would require such a data structure, and so would philosophers who hold that all knowledge must be explicitly represented in our minds. But there are two alternatives which would avoid the contradictions inherent in the information-processing model, by avoiding the idea that everything we know must be in the form of some explicit symbolic representation.

One response, shared by existential phenomenologists such as Merleau-Ponty and ordinary-language philosophers such as Wittgenstein, is to say that such "knowledge" of human interests and practices need not be represented at all. Just as it seems plausible that I can learn to swim by practicing until I develop the necessary patterns of responses, without representing my body and muscular movements in some data structure, so too what I "know" about the cultural practices which enable me to recognize and act in specific situations has been gradually acquired through training in which no one ever did or could, again on pain of regress, make explicit what was being learned.

Another possible account would allow a place for representations, at least in special cases where I have to stop and reflect, but would

stress that these are usually nonformal representations—more like images, by means of which I explore what I *am*, not what I *know*. We thus appeal to *concrete* representations (images or memories) based on our own experience, without having to make explicit the strict rules and their spelled out *ceteris paribus* conditions as required by *abstract* symbolic representations.

The idea that feelings, memories, and images *must* be the conscious tip of an unconscious frame-like data structure runs up against both *prima facie* evidence and the problem of explicating the *ceteris paribus* conditions. Moreover, the formalist assumption is not supported by one shred of scientific evidence from neurophysiology or psychology, or from the past "successes" of AI—whose repeated failures required appeal to the metaphysical assumption in the first place.

AI's current difficulties, moreover, become intelligible in the light of this alternative view. The proposed formal representation of the background of practices in symbolic descriptions, whether in terms of situation-free primitives or more sophisticated data structures whose building blocks can be descriptions of situations, would, indeed, look more and more complex and intractable if minds were not physical symbol systems. If belief structures are the result of abstraction from the concrete practical context, rather than the true building blocks of our world, it is no wonder the formalist finds himself stuck with the view that they are endlessly explicable. On my view, the organization of world knowledge provides the largest stumbling block to AI precisely because the programmer is forced to treat the world as an object, and our know-how as knowledge.

Looking back over the past ten years of AI research we might say that the basic point which has emerged is that *since intelligence must be situated it cannot be separated from the rest of human life*. The persistent denial of this seemingly obvious point cannot, however, be laid at the door of AI. It starts with Plato's separation of the intellect or rational soul from the body with its skills, emotions, and appetites. Aristotle continued this unlikely dichotomy when he separated the theoretical from the practical, and defined man as a *rational* animal—as if one could separate man's rationality from his animal needs and desires. If one thinks of the importance of the sensory-motor skills in the development of our ability to recognize and cope with objects, or of the role of needs and desires in structuring all social situations, or finally of the whole cultural background of human self-interpretation involved in our simply knowing how to pick out and use chairs, the idea that we

can simply ignore this know-how while formalizing our intellectual understanding as a complex system of facts and rules is highly implausible.

Great artists have always sensed the truth, stubbornly denied by both philosophers and technologists, that the basis of human intelligence cannot be isolated and explicitly understood. In *Moby Dick*, Melville writes of the tattooed savage, Queequeg, that he had "written out on his body a complete theory of the heavens and the earth, and a mystical treatise on the art of attaining truth; so that Queequeg in his own proper person was a riddle to unfold, a wondrous work in one volume; but whose mysteries not even he himself could read" (1851/1952, p. 477). Yeats puts it even more succinctly: "I have found what I wanted—to put it in a phrase I say, 'Man can embody the truth, but he cannot know it'."

Notes

1. This view is worked out further in Heidegger (1927/62); see especially p. 93 and all of section 18.
2. This is John Searle's way of formulating this important point. In a talk at the University of California at Berkeley (October 19, 1977), Schank agreed with Searle that to understand a visit to a restaurant, the computer needs more than a script; it needs to know everything that people know. He added that he is unhappy that as it stands his program cannot distinguish "degrees of weirdness". Indeed, for the program it is equally "weird" for the restaurant to be out of food as it is for the customer to respond by devouring the chef. Thus Schank seems to agree that without some understanding of degree of deviation from the norm, the program does not understand a story even when in that story events follow a completely normal stereotyped script. It follows that although scripts capture a necessary condition of everyday understanding, they do not provide a sufficient condition.
3. At the Society for Interdisciplinary Study of the Mind, Symposium for Philosophy and Computer Technology, State University College, New Paltz, NY, March 1977.

Minds, Brains, and Programs

7

John R. Searle
1980

What psychological and philosophical significance should we attach to recent efforts at computer simulations of human cognitive capacities? In answering this question I find it useful to distinguish what I will call "strong" AI from "weak" or "cautious" AI. According to weak AI, the principal value of the computer in the study of the mind is that it gives us a very powerful tool. For example, it enables us to formulate and test hypotheses in a more rigorous and precise fashion than before. But according to strong AI the computer is not merely a tool in the study of the mind; rather, the appropriately programmed computer really is a mind in the sense that computers given the right programs can be literally said to *understand* and have other cognitive states. And, according to strong AI, because the programmed computer has cognitive states, the programs are not mere tools that enable us to test psychological explanations; rather, the programs are themselves the explanations. I have no objection to the claims of weak AI, at least as far as this article is concerned. My discussion here will be directed to the claims I have defined as strong AI, specifically the claim that the appropriately programmed computer literally has cognitive states and that the programs thereby explain human cognition. When I refer to AI, it is the strong version as expressed by these two claims which I have in mind.

I will consider the work of Roger Schank and his colleagues at Yale (see, for instance, Schank and Abelson 1977), because I am more familiar with it than I am with any similar claims, and because it provides a clear example of the sort of work I wish to examine. But nothing that follows depends upon the details of Schank's programs. The same arguments would apply to Winograd's (1972) SHRDLU, Weizenbaum's (1965) ELIZA, and indeed, any Turing-machine simulation of human mental phenomena.

Briefly, and leaving out the various details, one can describe Schank's program as follows: the aim of the program is to simulate the