CAN A SYSTEM BE INTELLIGENT IF IT NEVER GIVES A DAMN?

Thomas Edelson

Department of Computer Science
Georgetown University
Washington, DC 20057

ABSTRACT

I explore whether all types of cognitive faculties are possible in a system which does not also have "affective faculties" such as motives and emotions. Attention is focused on the human cognitive faculty of understanding the human affective faculties, and it is suggested that we do this in part by using ourselves as models of other people. Therefore any system which could perform as well as we do on this task would incorporate much knowledge about human affective faculties, and would probably have embedded within it a "model human" which would possess affective faculties. However, this would not necessarily manifest itself directly in the embedding system's behavior; thus the embedding system, unlike the embedded one, need not have affective faculties. That's a relief, because we might prefer that it didn't.

Keywords: emotions; feelings; metacognition; motives; common-sense psychology; folk psychology; naive psychology; Haugeland, John.

1. Introduction

Can there be thinking without caring, reason without emotion, intelligence without engagement? I will argue that:

(1) There are some types of thinking which probably can't be separated from caring, in the sense that we won't know how to build systems that can do those kinds of thinking until we know how to build systems that care.

(2) However, saying that we will need to know how to build caring systems does not imply that we must actually do so; it is possible to build thinking systems which do not care.

(3) If I'm right about (1), then people interested in practical uses of AI had better hope that I'm right about (2) as well. The alternative, that certain kinds of intelligence are possible only in caring systems, is unattractive, because we might not want applied AI systems to be caring systems.

Haugeland (1985) seems to agree with my claim (1); indeed, much of his last chapter can be read as an argument for it. However, he disagrees with claim (2), and does not discuss (3).

In this section I will clarify these claims; then I will turn to arguing for them one by one.

I will drop the informal term "caring" and speak instead of "engaged systems" and systems which have "affective faculties".

An engaged system is one for which it makes sense to talk about its motives. In explaining what it does, we find it natural to say things like "it did x because it wanted to accomplish y". Furthermore, not all these motives are simply subgoals on the way to accomplishing some one thing -- they are not all means to a single end. Rather, the system is capable of having multiple, independent motives, each of them an "end in itself" rather than merely a means.

Also, it is capable of change over time in its set of motives. It can gain new motives and lose old ones. A corollary of this is that observers will need to make inferences from its behavior in order to figure out what its motives are at any particular time.

It should be clear that human beings are engaged systems. I hope it is also clear that almost no computer-based systems, so far at least, are engaged systems (nor are they intended to be). We may say of a question-answering system that it "wants" to figure out the correct answer, of a chess-playing program that it "wants" to win (Haugeland, 1985), of an operating system that it "wants" to carry out our commands. But these usages are metaphors, not literal statements; such systems do not have the complexity of motives required for an engaged system.* They acquire new motives only as intermediate goals, or because we assign them.

I have defined an "engaged system" as one with multiple, independent, changeable motives. What about "affective faculties"? This is intended to be a catchall term for everything in the mental domain which is not strictly rational or cognitive. Thus belief and thinking are not affective, but motives and emotions are.

* A system with one motive doesn't really have any motives. Compare Haugeland on "semantic intrigue" (pp. 216-217).

Sloman and Croucher (1981) argue convincingly that (what I call) an engaged system will also have other affective faculties, such as emotions. I accept this conclusion; furthermore I note that their argument equally supports the converse claim, that a system with emotions must be an engaged system. Therefore I will use "engaged system" and "system with affective faculties" almost interchangeably.

2. Cognition About Affect

In this section I argue for the claim that any reasonably complete model of human cognitive faculties will include a model of human affective faculties: the two domains cannot be separated.

In an excessively simple form, the argument is: any reasonably complete model of human cognitive faculties must include the ability to reason about human behavior and the faculties which produce it, including the affective faculties; for a computer system to do this, it will need to include a model of the affective faculties.

The problem with the argument as stated is that it applies to too many things. One could equally well say:

"Any reasonably complete model of human cognitive faculties must include the ability to reason about the weather and the processes which produce it; for a computer system to do this, it will need to include a model of the global atmosphere. The cognitive domain cannot be separated from the meteorological domain."

If the cognitive domain is inseparable from the affective domain in only the same way that it is inseparable from the meteorological domain, that's not such a big deal. It's just another way of saying that understanding of human affect and motives -- more generally, understanding of "human nature" -- is part of the "common-sense knowledge" which AI systems of the future are going to need (Lenat et al, 1986).

But I claim that our knowledge about "human nature", including the affective faculties, is significantly different from other kinds of common-sense knowledge. I want to suggest that when humans perform cognitive tasks which happen to involve thinking about human affect, then something special is going on. In thinking about humans, we take special advantage of the fact that we are humans. .Therefore, to reproduce this particular set of cognitive capacities in a system which is not itself human will present a special (though not necessarily insuperable) problem for AI.

What is this special process which (I claim) is involved in thinking about people? It consists of using oneself as a model of the person one is thinking about. I will now ask you, the reader, to provide your own example of this process, by acting as an experimental subject. All you have to do is to answer a question about how someone would feel in a fictional situation.

Imagine an engineer called in by her boss after a space shuttle accident. This engineer had argued that it was not safe to launch the shuttle, and it now appears that her doubts were well-founded. The manager asks the engineer not to reveal her pre-launch objections; if asked by the board of inquiry, to say that she had fully concurred with the recommendation to launch. The manager adds that the company will not forget who was, and who was not, a team player at this critical time.

How does the engineer feel? I hope you agree that most people in such a situation would feel dismay, anger, and fear. But how do you know that? Is it an inductive inference from similar situations which you have seen people in, or been in yourself? It could be that; it probably is partly that.

But I would suggest that it is also something else. In answering a question about hypothetical people in a hypothetical situation, we imagine ourselves into that situation. And "imagining" here is not a purely cognitive process. It doesn't just involve remembering how our (or others') affective faculties functioned in the past; it involves our affective faculties now. If you imagine yourself into a situation like that of the engineer, you may actually feel her reactions yourself.

I am proposing that the process of imagining a situation literally activates some of the same brain processes which occur in response to a real situation; somehow we are able to "fake" the neural messages which would have come from our sensory apparatus. As a result, at least part of our brain responds exactly as it would in the real situation (for indeed, since the messages it is receiving are the same, it can't tell the difference). Therefore the internal process which normally corresponds to getting angry (for example) may actually occur. We observe this going on in ourselves, conclude that it would happen in the imagined situation, and report accordingly.

What I am offering is a piece of speculative psychological theory. I have been unable to find any mention of this theory in the psychological literature. There are, however, interesting parallels between what I am suggesting and some work on mental images; for example (Finke, 1986).

3. A Computational Analogy

In this section I will try to make my psychological hypothesis more concrete, by suggesting an analogy between it and a computational process.

I suggest that what goes on in a person who is imagining himself in a situation is something like what goes on in a "virtual machine" facility such as IBM's VM/370 (Buzen and Gagliardi, 1973; IBM, 1972). A virtual machine is an imaginary computer being simulated by a real one; the VM/370 control program feeds inputs to the virtual machine in such a manner that these inputs appear to be coming directly from real input devices. Sometimes they are in fact coming from devices of the same general type, but sometimes not: when a virtual machine reads from its "card reader" it may be receiving data which in fact has never been on a card, but was instead generated by some other virtual machine.

Similarly, when a virtual machine is ready to produce output, it executes a Start I/O instruction, which would, in a real machine, cause an I/O channel to begin the output operation. Since this is a virtual machine, however, the output request is intercepted by the control program, which then simulates the operation. What is actually done with the data may be quite different from what the virtual machine "thinks" it is doing.

The analogy is imperfect: for one thing, an unprogrammed 370, real or virtual, has no affective faculties (and no human-like cognitive faculties, either). For another thing, the purpose of using a virtual machine is usually not to try to figure out what another machine would do. Nevertheless, the VM/370 analogy is particularly apt in at least one way: the program in the virtual machine is not in fact merely simulated, in the sense of being carried out by a software interpreter. It is executed directly by the hardware (or microcode) of the real 370. Only the inputs and outputs are simulated or redirected.

In a similar way, I propose, a human being may generate inputs (representing an imagined situation) which don't come directly from the outside world, but which are fed to affective faculties just as if they had; these faculties generate outputs which would normally lead to action, but which are trapped and redirected (fed as inputs back into the part of the brain which is managing the simulation experiment).

Since the virtual machine's program is executed directly by the real machine, the VM/370 software system need not, and does not, include a software interpreter for the 370 instruction set. Similarly, if human beings use the mechanism I am suggesting, then they can "model" the affective faculties of other people (or themselves), without having a set of mental representations of how those affective faculties work. I don't have to know what the causal chain is that leads from sensory inputs to motor outputs; I can invoke that causal chain by providing bogus sensory input, and trap and observe the resulting output signals, without being able to describe what goes on in between.

If I am right, there is indeed a significant difference between common-sense knowledge about people and common-sense knowledge about weather. The former, unlike the latter, may in part not be "knowledge" in the usual sense at all, that is, information encoded somehow in memory. And we didn't have to learn it.

We certainly don't rely only on this method. We also do have and use ordinary, explicit knowledge about people. The "virtual person" method can be used when, lacking any more specific information, we go on the assumption that others would act the same way we would. When we have information from experience about how a specific person acts in a given kind of situation, we are likely to get a better prediction by using that information than we could get by putting ourselves in her place.

4. How to Model Affect

If my speculation about how humans understand each other is correct, how then can a computer system be programmed to equal human performance at the task of understanding humans? There seem to be two possible approaches.

The first approach is to build an engaged system: one which has complex multiple motives, feelings, and the rest of the affective faculties. Then it can do the "virtual person" trick to understand people by analogy with itself.

The other approach is to build a system which does not share the affective faculties, but reasons about them in the same way that it reasons about any other topic. This would require giving it a knowledge base about human nature which is more complete than the one which actual humans have. (Humans do have explicit knowledge about human nature, but the machine would need more of it to make up for its lack of non-explicit "knowledge".)

It may seem that the two approaches would lead to quite different kinds of models of the affective faculties. However, I want to argue in this section that they probably would not; that if the two approaches are both successfully pursued, they will probably produce quite similar models, differing only in the way that the models are connected to the outside world.

Why would we initially expect the models to be different? Because designers following one approach would be most naturally led to look at the affective faculties "from the inside", while those following the other approach would look at them "from the outside". If you are trying to build a system which actually has human-like affective faculties, then you learn what you can about the way those faculties work in human beings, and build a system which works the same way. By contrast, if your objective is merely to give your system a theory about human nature which it can use to make predictions and answer questions, then it is most natural to think of the system as simply observing human beings from the outside, treating them as black boxes; either the system or the system builder would then need to come up with a theory which predicts the outputs of those black boxes, given their inputs.

However, the important question is what works; and it may turn out that the "black box" approach just doesn't work well enough. What this approach amounts to is trying to determine a function by looking at its behavior over some large but finite set of input values. If the function is allowed to have arbitrary complexity, then any such set of data underdetermines what the function is. And therefore mistakes may be made in predicting its future behavior.

The goal, of course, is not perfect prediction but merely prediction as good as that achieved by humans (which certainly isn't perfect). But if, as I hypothesize, humans use the "virtual person" approach, which gives them access to the inside of the box, it may turn out that this enables them to do better on some predictive tasks than the black box approach ever could.

Also, even if the black box approach could work in theory, it may still be easier in practice to construct a working predictor by taking advantage of knowledge about the inner mechanisms of human affective faculties.

It is possible that the black box approach will, even in practice, produce acceptable performance. If so, it could still turn out that the model constructed by this approach is similar to the one created from knowledge of the inner mechanisms. Obervation from the outside may lead to a computational model which is, in fact, similar in structure (perhaps even mathematically isomorphic) to one created the other way, so that the black box modeler will have modeled the inner mechanisms without knowing it.

Admittedly, there is the residual possibility that the black box approach will work, and produce a model quite different from what you'd get if you worked with knowledge of the inner mechanisms. I can't rule that out; time will tell.

5. Cognition Without Affect? Yes

I have now completed argument for my first claim. That argument may be summarized as follows:

Any reasonably complete simulation of the human cognitive faculties must include the human ability to reason about humans, including their affective faculties; thus it will include a model of their affective faculties. Humans probably do such reasoning, in part, by using themselves as models of the other (what I have called the "virtual person" technique), thus having and using access to the inner mechanisms of the affective faculties, rather than only observing them from the outside. It is probable, though not certain, that any system which matched our performance on these tasks would do it by having a model which also was based on, or at least unintentionally resembled, these inner mechanisms of the affective faculties. Therefore, in designing such a system, the designers would probably have learned so much about the affective faculties that they would be able to build a system which had affective faculties of its own -- an engaged system -- if they wished to do so.

Now I will argue for the second claim: that it is nevertheless possible to build a system which can reason as well as we do about human affective faculties, but which does not itself share them.

How do we tell whether a system has affective faculties? By observing its behavior. Here's a rather simplified example:

Suppose we tell a system that we don't think its performance is as good as some other system's. Suppose it replies, "If you think that one is better, why don't you use it?" Suppose, more importantly, that it then refuses to answer any questions for us until we retract our statement. Then we would be inclined to say that this system has a self-image that it cares about, and that it is capable of anger.

Given such criteria for what counts as an engaged system, does any system which contains a model of an engaged system necessarily have to be one itself? Clearly not. The question is not how lifelike the model is internally, but how it is connected to the system as a whole.

For example, contrast the system described above, which is sensitive to criticism (call it Fred) with one that simply understands how humans can be sensitive to criticism (call it Flo). If you tell Flo a story about a person who is criticized and then refuses to cooperate, Flo can explain why this occured. How? Very likely because Flo contains a model (call it Moe) which has, as humans have, the tendency to be sensitive to criticism. Flo finds out how a person would react to a critical remark by making that remark internally to Moe and seeing how Moe responds. But Moe's response is not passed directly to us; it is only used by Flo to draw a conclusion.

If we make a critical remark to Flo itself, it doesn't respond as either Fred or Moe would. Perhaps it ignores the input entirely, judging it irrelevant to its programmed goals; perhaps it takes our word for the statement and stores it away in its knowledge base; but it does not retaliate or show any other general disturbance in its question-answering behavior.

Therefore Flo is not an engaged system, though it contains a model of one.*

Haugeland (1985), in a section titled "Wouldn't a Theory Suffice?", uses essentially the same example to argue for a conclusion opposite to the

* To refer back to the analogy with virtual machines, Flo is not like a VM/370 system, because here the virtual machine (Moe) has a different architecture from that of the real machine in which it is embedded (Flo).

one I have just argued for.** Agreeing that Flo has no affective faculties, he wants to conclude that Flo doesn't really understand human affective faculties either -- and thus, more generally, a system which didn't have the faculties couldn't understand them. The crux of his argument is as follows:

"...Flo has to give Moe the whole story and then simply plagiarize his response. In other words, the idea of Flo understanding the stories cognitively, with occasional reference to her theory of selves, is a total scam. Moe (in whom there is no such segregation [between cognition and affect]) is doing all the work, while Flo contributes nothing."

But Moe is part of Flo! Haugeland's argument implicitly rejects behavioral criteria for determining whether Flo understands or not, though Haugeland accepts such criteria elsewhere in the book (see p. 215). If we ask Flo questions and get answers which seem to indicate understanding, then Flo understands, no matter if she consults a whole platoon of inner specialists in the process.

To use Haugeland's own terminology, in this passage he abruptly and temporarily abandons his otherwise preferred stance as a "skeptic" about AI, and speaks instead as a "debunker" -- just like (Searle, 1980). In effect he is saying: "It may look like understanding from the outside, but when you see the way it works inside, you'll realize that it isn't really understanding after all."

I conclude that cognition and affect can be separated -- or more precisely, decoupled. A system which can understand the affective faculties needs to contain a model of them; nevertheless, it is possible to arrange matters so that these affective faculties do not directly manifest themselves in the personality which the system presents to the outside world.

But when I say "it's possible", I don't mean to imply that it's just a matter of rolling up our sleeves and doing it. All I really mean is that it hasn't yet been convincingly been shown not to be possible. There are serious unsolved problems which face any attempt to create a system which can really understand human affect -- whether or not you want it to share those qualities as well as understand them. The attempt may yet be abandoned as infeasible, or just not worth the enormous effort -- if not as impossible in principle.

6. What Do We Want?

So far as we can tell, then, building a system which understands human affective faculties without sharing them is just as feasible (or infeasible) as building one which does share them. Given the choice, which type of system would we rather have?

Someone will want to try to build an engaged system, at least for research purposes: to test theories about how humans work -- and, of course, simply to show that it can be done.

What about applied AI? There, I suggest, one would usually rather not have a system which displayed human-like affective faculties. Do you want to have to worry about the possibility that your expert system will refuse to help you, because you have insulted it? Which would you rather deal with, Fred or Flo?

The choice is up to you.

REFERENCES

Buzen, J. P., and U. O. Gagliardi. "The Evolution of Virtual Machine Architecture". In Proc. National Computer Conference, New York, June, 1973, pp. 291-299.

Finke, R. "Mental Imagery and the Visual System". Scientific American 254:3 (March 1986) 88-95.

Haugeland, J. Artificial Intelligence: The Very Idea. Cambridge, Massachusetts: MIT Press, 1985.

IBM Corporation. IBM Virtual Machine Facility/370 -- Planning Guide. Pub. #GC20-1801-0, 1972.

Lenat, D., M. Prakash, and M. Shepherd. "CYC: Using Common Sense Knowledge to Overcome Brittleness and Knowledge Acquisition Bottlenecks". AI Magazine 6:4 (Winter 1986) 65-85.

Searle, J. "Minds, Brains, and Programs". Behavioral and Brain Sciences 3:3 (September 1980) 417-424.

Sloman, A., and M. Croucher. "Why Robots Will Have Emotions". In Proc. IJCAI-81, Vancouver, British Columbia, August, 1981, pp. 197-202.

** See pp. 242-243. I have borrowed the names.