AI Magazine Volume 11 Number 3 (1990) (© AAAI)

An Essay Concerning Robotic Understanding[1]

Jerome Feldman

The question of whether a computer can think like a person is once again a hot topic. Somewhat to my surprise, this philosophical question seems to have direct practical implications for AI, especially language understanding. The following analysis has been helpful to me and might be of some value to others.

When we use words such as think, understand, and wish, we typically refer to the human experience of these activities. When I want to emphasize this point, I use the notation x/h for human. However, we can also say that a computer system knows something in a data processing sense. We could say that a database (or an apolitical person) knows/d that Vilnius is the capital of Lithuania without having any idea of what this fact means to many people. The apolitical person could be given a little history and with his(her) knowledge of nationalism come to understand/h fully. Although no current program can exploit such general knowledge, one can imagine computer systems that would capture the data processing aspect of this enlightenment, but what about the ecstatic feeling of being in a crowd that is surging through the streets?

The basic question seems to be, Is there a fundamental difference between understanding/d and understanding/h, and if so, does it matter for practical AI? Everyone agrees that emotions, aspirations, and so on, are a crucial aspect of our mental life and that no superficial realization of x/d will capture x/h. If Smith in the Chinese Room received tiles saying he had won the lottery or that the building was on fire, no output of tiles would be at all the same as his behavior on hearing such statements in English. To survive the fire, the Chinese Room as a system would need to connect (ground) the symbols to its physical properties and abilities. We also have behavioral and physiological evidence that strong emotional responses affect human data processing in fundamental ways. The deep interconnections among mind and body are the crux of the issue.

Two basic lines of reasoning are used to support the notion that computers or robots eventually might fully achieve x/h. The more common and less interesting argument is based on ever more detailed simulation. To take the simulation story to the extreme, suppose a computer system simulated every molecule in the brain of some idealized person; how could it not have x/h? I believe that there would still be detectable differences, but it doesn't much matter because there are plain and fancy reasons why such a simulation is impossible. Because the biochemistry of drugs, hormones, and neurotransmitters plays a central role in human information processing, it is unlikely that a coarser simulation will automatically capture x/h.

The other way by which we might have computers think like humans is less direct and requires a longer story. Much of x/h concerns the human body, its homeostasis, and its interactions with the world. Thus, we assume our aspiring computer will need interior and exterior senses and an ability to interact with the world; that is, it is a robot. We could (even now) endow this robot with programs that can interpret internal sense readings (low battery, wheel slippage, and so on) as being good or bad for the robot along various dimensions. Such a robot could come to correlate wet pavements with slippery wheels and legitimately issue the statement, "I don't like to go out when it's wet." This story linking the robot's decisions to the external world is an informal example of what I believe is a critical development: a data processing definition of (intrinsic) intensionality. The robot has its own goals/d and can learn/d which combinations of sense readings and actions further these goals. Other intensionality words, including consciousness/d can be similarly treated. Now, different robots might have different goals/d and, consequently, beliefs/d, desires/d, and so on. (There is an evolutionary version of this story, but it isn't relevant here.) For our purposes, the goal is to make robots that are as human-like as possible. Suppose we use x/r to denote the use of intensionality words with respect to these humanoid robots. Now the question becomes, Could we develop these systems to the point where x/h and x/r were used interchangeably. In this case, we would mean exactly the same thing when we said that Mary or R2D2 understands Proust or loves John.

To explore the question of whether x/r could equal x/h, we must look more closely at x/h, particularly at understand/h. We actually use understand/h loosely, normally excluding infants, idiots, and so on. We acknowledge that there are strong limitations on the extent to which we can convey understanding/h across barriers such as gender, age, race, and culture. There are understandings/h that we share with our colleagues and not with our family and vice versa. If we built an expert system that cared/r how often and successfully it ran, it could well turn out that this system and an expert person could share deeper understanding (and beliefs and desires) within this domain than the person could with most other people. An analogous situation would be a champion horse-and-rider team.

Nevertheless, there is a basic sense in which understand/h (and x/h in general) does refer to our shared human experience, and human experience is based in the human body, brain, biochemistry, and so on. It seems possible (to me, almost certain) that robots that are physically very different from people will in general have x/r that is different from x/h. This difference does not depend on phenomenology; two robots with radically different sensors and mechanisms would find it hard to communicate. If the basis in the body is correct, then it is critically important if one also accepts (following linguistic evidence) the bodily grounding of semantics of language. The notion here is that many of the most basic components of natural language are directly grounded in our sensory and motor apparatus. The bodily grounding hypothesis is obvious for words such as see, want, and push but extends to encompass notions of space, forces, and so on. The strong form of the bodily grounding hypothesis is that much of the

rest of language is interpreted by mappings to this core. To the extent that this is true, robots will find/r it hard to communicate with people. People routinely invent new language usage and are usually understood without elaboration. Meaning based on the body provides an explanation. Conceptual extensions that automatically occur in humans would be a mystery to robots with radically different bodies. Of course, making the robot connectionist wouldn't help.

The take-home lesson for me is the following: A presumption of shared experience is the basis for communication. If we want computer systems to understand/h (or learn/d) natural language well enough to meet AI goals, we need to explicitly account for the x/h aspects that underlie much of language/h and thought/h. We could try to build robots that can understand/r human experience by building robots and pushing their x/r to work as much as possible like x/h. Although interesting and fun, this solution is not likely to work in the short term. The alternative is to explicitly view the problem as one of communicating among alien species. Our programs should try to incorporate as much knowledge/d of human x/h as needed for the tasks involved. The common way to attempt this incorporation is to include lots of rules about human x/h. The previous analysis suggests that rules about human experience will never be adequate, and we must work on simulations of human understanding. For example, one should not try to list all the conditions that might cause dizziness but rather include a vestibular model good enough for prediction. Connectionist techniques appear to be required because they make it possible to capture the evidential, situational, multifaceted character of human thought.

However, unfortunately, even the simplest natural language domains (such as scenes of circles and squares) entail a great deal of knowledge/h to understand/d all that people might want to communicate. To the extent that we fail to adequately capture x/h, we should have greatly reduced expectations of our programs as teachers, therapists, judges and of any application where the richness of human experience is important.

NOTES

1. "An Essay Concerning Human Understanding" was published in 1690 by John Locke.