Toward Human-Level Cognitive Adequacy

Our long-range aim is to design and implement common sense in a computer. Common sense in a computer is a bit hard to define, but the idea we are aiming at is comparable to human-level common sense (often understood as distinct from expert or special cleverness). For instance, solving the mutilated checkerboard problem takes a special clever insight, and thus is *not* what we have in mind. But note that much of the AI community would not draw the definitional lines as we have; for that reason, we sometimes use the expression "cognitive adequacy" to refer to our conception. This is intended to suggest a kind of general-purpose reasoning ability that will serve the agent to "get along" (learning as it goes) in a wide and unpredicted range of environments. Consequently, one hallmark of common sense is the ability to recognize, and initiate appropriate responses to, novelty, error, and confusion. Examples of such responses include learning from mistakes, aligning action with reasoning and vice versa, and seeking (and taking) advice.

A closely related hallmark is the ability to reason about anything whatever (that is brought to one's attention). This does not mean being clever about it, or knowing much about it, or being able to draw significant conclusions; it can mean as little as realizing that the topic is not understood, asking for more information, and learning appropriately from whatever advice is given. That may seem like very little, if our model is to have clever solutions to tricky problems. But consider this: virtually no AI programs exhibit even that "little" amount of elementary common sense; they are not able to know when they are confused, let alone seek – and use – clarifying data. On the other hand, cleverness – in highly limited domains and for tightly specified representations – has been built into many programs, a kind of "idiot savantry" that fails utterly when outside those narrow strictures.

One large piece of what is needed for cognitive adequacy, then, is what we call "perturbation tolerance": the ability to keep going adequately when subjected to unanticipated changes. This includes changes to the knowledge base (KB); e.g. the changes might introduce inconsistencies, or make a goal impossible or ambiguous. Worse, the knowledge representation (KR) system might change (new terms, new meanings for old terms, different notational conventions, etc), especially if other agents are involved; and of course there are typos (missing parentheses and the like) that appear to defy any prearranged methodology. And there are changes to physical sensors and effectors, and how things in the world work.

Then what is it to "keep going adequately" in the face of such changes? Among other things, this will require (i) never "hanging" or "breaking"; (ii) recognizing when there is a difficulty to be addressed; (iii) making an assessment of options to deal with the difficulty; (iv) choosing and putting an option into action. Such a suite of abilities will require keeping track of one's own history of activity, including one's own past reasoning. Such an agent will then, in Nilsson's phrase, have a *lifetime of its own*, and keeping track of its own processes and history will allow it to look back at what it is doing and use that knowledge to guide its upcoming behavior.

People tend to do this well. We think there is strong evidence as to how, and we have a specific hypothesis about it and how to build it in a computer. In a nutshell, we propose what we call the *metacognitive loop*—that allows a reasoning agent to note, and attempt to correct, errors in real time—as the essential distinguishing feature of cognitive adequacy, including perturbation tolerance. And we claim that the state of the art is very nearly where it needs to be, to allow this to be designed and implemented.