What's Wrong with Artificial Intelligence

Rich Sutton

11/12/2001

I hold that AI has gone astray by neglecting its essential objective --- the turning over of responsibility for the decision-making and organization of the AI system to the AI system itself. It has become an accepted, indeed lauded, form of success in the field to exhibit a complex system that works well primarily because of some insight the designers have had into solving a particular problem. This is part of an anti-theoretic, or "engineering stance", that considers itself open to any way of solving a problem. But whatever the merits of this approach as engineering, it is not really addressing the objective of AI. For AI it is not enough merely to achieve a better system; it matters how the system was made. The reason it matters can ultimately be considered a practical one, one of scaling. An AI system too reliant on manual tuning, for example, will not be able to scale past what can be held in the heads of a few programmers. This, it seems to me, is essentially the situation we are in today in AI. Our AI systems are limited because we have failed to turn over responsibility for them to them.

Please forgive me for this which must seem a rather broad and vague criticism of AI. One way to proceed would be to detail the criticism with regard to more specific subfields or subparts of AI. But rather than narrowing the scope, let us first try to go the other way. Let us try to talk in general about the longer-term goals of AI which we can share and agree on. In broadest outlines, I think we all envision systems which can ultimately incorporate large amounts of world knowledge. This means knowing things like how to move around, what a bagel looks like, that people have feet, etc. And knowing these things just means that they can be combined flexibly, in a variety of combinations, to achieve whatever are the goals of the AI. If hungry, for example, perhaps the AI can combine its bagel recognizer with its movement knowledge, in some sense, so as to approach and consume the bagel. This is a cartoon view of AI -- as knowledge plus its flexible combination -- but it suffices as a good place to start. Note that it already places us beyond the goals of a pure performance system. We seek knowledge that can be used flexibly, i.e., in several different ways, and at least somewhat independently of its expected initial use.

With respect to this cartoon view of AI, my concern is simply with ensuring the correctness of the AI's knowledge. There is a lot of knowledge, and inevitably some of it will be incorrrect. Who is responsible for maintaining correctness, people or the machine? I think we would all agree that, as much as possible, we would like the AI system to somehow maintain its own knowledge, thus relieving us of a major burden. But it is hard to see how this might be done; easier to simply fix the knowledge ourselves. This is where we are today.