What's Wrong With HAL?

<Extended abstract>

Ken Ford
NASA-Ames Research Center

Pat Hayes
Institute for Human & Machine Cognition
University of West Florida

Aliens in our midst

While playing against Deep Blue, Kasparov reported feeling an 'alien intelligence' across the board: something that was undoubtedly thinking, but not in the way that any human thinker would. Kasparov feels he is defending the honor of the human race. If cloning a single sheep can put the leaders of the Western world into such a tizzy, one wonders how will they react to the prospect of inhuman mechanical intelligences in our midst.

We will argue here that the intelligences that AI is likely to make will indeed be 'alien', but rather than look on this with dread, we should rejoice in this alienness. The last thing AI should seek to do is to make machines which think exactly in the way that we humans think.

Sorry, Dave, I can't do that

The problem can be illustrated very well with the fictional robot HAL in Kubrick's movie "2001". HAL was intelligent, no doubt, and in some respects was made to be as human-like as possible. HAL was an expert thinker on many subjects, but it was also designed in part to respond socially, acting as a rather special crew member. Rather than being simply a source of expertise upon possible difficulties of social interaction, HAL was designed with an internal architecture which could support empathy and other human-like emotional reactions, to enable it to actively participate in a small human society.

While having to seem honest, HAL was in fact entrusted with dangerous secrets, and was obliged to lie to its friends. When Dave caught HAL in a small lie, a seed of doubt and mutual suspicion began to disturb the serenity of the enclosed society within the spacecraft, leading rapidly to an atmosphere of mutual supicion which was sufficiently intense to push HAL into psychosis, and the computer began to suffer from paranoid delusions.

The point of this is to emphasise how much HAL's deviant behavior resulted from a breakdown in its social relationships with the crew. This could only happen because HAL was constructed to imitate human thinking in a remarkably thorough way. HAL was built to be *too human*.

Chess: Man vs. Machine?

One criticism of chess-playing programs often made within AI is that since chess is an artificial, restricted, 'brittle' domain, success at chessplaying tells us little about general techniques for learning how to improve problem-solving skills (say). A rather different criticism made from outside AI is that regardless of who wins, human chess-players are still uniquely human because the machines think differently.

These are really two sides of the same objection, motivated by the legacy of Turing's old Test. Turing's motivation was to replace a vague goal of building an 'intelligent' machine by a more concrete one which could be tested objectively. Thinking of the task of building a chess-playing program in the same way renders both of these objections pointless. Nobody criticises a cheesegrater for being ill-adapted to making omelettes, or the fact that its methods don't yield insights into the general nature of friction, so why are similar objections thought to be relevant when applied to chess-playing programs or any other AI program designed to perform a particular intellectual task? The reason seems to be that a chess-playing program is thought of as merely a step towards a larger goal of making something which thinks like a human. From this perspective these criticisms make sense; but if ones goal is to make a chessplayer, having it think like a human is a positive disadvantage. While humans have many intellectual talents, they also have many weaknesses: limited short-term memory capacity, easily distracted by irrelevant stimuli, emotional reactions which divert energy and so on. The aim of building a chess-playing tool is to surpass human chess performance, not to imitate it.

Calling Deep Blue a 'tool' reflects an attitude towards AI which we think is a more promising one than the traditional Turing vision. The goal of AI, on this view, is to make machines which amplify and extend our intelligence, just as the steam shovel amplified and extended the way that our muscles can dig dirt. Another way to think of this chess match is as not being between Kasparov and Deep Blue, but between the current chess champion and a human being who is using Deep Blue to help him choose his moves. The remarkable thing is that the man with the chess-tool is able to do so well when he knows so little chess. Evidently this chess-tool is enormously useful, if chess playing is your business.

Deep Blue is much more like a tool than an independent agent. It doesn't have its own agenda. It has no purpose other that for which it was designed, and is highly optimised to this very specialised use. All it does is figure out a good move from a chess position. If we regard a scientific calculator as a tool, Deep Blue surely is in the same category. However, this is not to say that Deep Blue is not intelligent. If chess-playing requires

intellectual ability, then of course a successful chess-playing program is intelligent. Suppose someone had a dog that was the second-best chess player in the world; everyone would surely agree that this was a very smart dog, even if it couldn't hold a human conversation or fully understand the concept of 'chess match'. The lesson of AI is that we can make intellectual tools — cognitive prostheses — which have no independent social role, no human pretensions or weaknesses, but yet have genuine intellectual power. They really do think, in many ways better than we can. Which is, of course, our reason to build them. Like any other tool, they help us do things we could not manage to do without them.

Notice that the alien quality of a tool is critical. Were it not alien in some way, there would be no use for it. That is what tools are for: to perform some function which is beyond our capacity to perform without it. Whether that task be loading coal or playing chess, the relationship between human and machine is similar. Mechanizing these talents beyond that of a naked human being does not reduce or demean human dignity or threaten us, but rather expands our human capability. Before industrial mechanization, a strong muscular body was a valuable economic resource. Strong bodies are still admired, but now only for essentially aesthetic reasons. Until recently it took years of dedication to learn to play chess really well, but now for less than a hundred dollars anyone can buy a chess tool and use it to play chess close to the master level.

However it is worth noting that, like many AI programs, Deep Blue is not *totally* alien. Human chess players also perform lookahead and evaluate future positions. The differences between Deep Blue and Kasparov are more quantitative than qualitative. Our second critic is right to claim that human thinkers have talents which we cannot mechanize (yet), but the moral of these chess matches is that these other skills which Kasparov uses — the ability to recognize 'meaningful' chess patterns, for example — are rendered unnecessary simply by magnifying a limited sub-collection of human talents by the speed and reliability of the computer. A sufficiently fast 'dumb lookahead' can overcome the craftiest human combinations of multiple mental skills. Once again, we urge an optimistic conclusion. Consider the possibility for making more useful synergies between man and computer when the computer is used to amplify and accelerate parts of *human* thought many orders of magnitude past the capacity of an unaided human brain.

On measurement

Finally, another reason why chess-playing programs are something of a model for AI is that chess performance can be measured quite objectively and qualitatively, in notable contrast to most of the performance goals which AI has tackled. We would argue that AI research generally should seek more such objective performance scales.