Agents, Decisions, Beliefs, Preferences, Science and Politics

David Poole
Department of Computer Science
University of British Columbia
Vancouver, B.C., Canada V6T 1Z4
poole@cs.ubc.ca
http://www.cs.ubc.ca/spider/poole/

August 10, 2006

Abstract

AI is about building intelligent agents; in particular, it is about deciding what an agent should do. What an agent should do depends on its capabilities, its beliefs and its preferences. AI has developed many techniques for deciding what to do. For a long time there was a tradition of developing rich representations without uncertainty, but now it is the time we need to develop rich representations that take uncertainty into account. I sketch a vision where computers will be able to find the best actions leaving decision makers to argue about values.

1 Decisions and Representations

AI is about building intelligent agents. What an agent should do depends on

- its capabilities; what it can do, including its computational limitation
- its beliefs; these are made from an agent's observations and what it has remembered
- its preferences; what it thinks is important, including its goals, but when it cannot achieve its goal with certainty, it needs to be able to trade off the outcomes of its actions

This is complete in the sense that two agents with the same capabilities, beliefs and preferences should do the same thing (even if there actions are stochastic).

There has been a long tradition of considering the problem of making decisions in economics, control engineering, psychology, etc., where the dominant paradigm is Bayesian decision theory and its multi-agent counterpart, game theory (with the only real challenges coming from psychology [Tversky and Kahneman, 1974]). For a long time Bayesian decision theory was rejected in AI, as probabilities were "epistemologically inadequate" [McCarthy and Hayes, 1969]. Indeed if you only consider probability over simple events or over static random variables, then the representation is inadequate. But probability does not have to be over simple events.

It was probably fortunate that AI researchers considered the case without uncertainty first. It is always simpler to state what is true in a representation than to state probability distributions over what is true in that representation. By studying representations we have been able to determine what was needed in a representation of the world. There are many sophisticated representations that have been developed without considering of the added complexity of decision-theoretic and probabilistic reasoning over these representations [Davis, 1990; Shanahan, 1997; Sowa, 2000]. We may not have been able to develop these representations if we were simultaneously worrying about probability distributions and preferences over these representations. But now is the time we need to consider rich representations with uncertainty and preferences.

2 A Vision

There are many visions of AI future in terms of robots, ubiquitous computers and the semantic web. These are typically about capabilities, but I will concentrate on beliefs and preferences.

Belief are about predictions, which is the role of science and machine learning. Agents will be able to make scientific theories about the world they will interact with, both the physical world and the social world. These theories will be tested by data and by acting in the world. We will determine which theories are best by considering their performance.

Preferences are about what is right and wrong, what is good or bad. Science or machine learning won't tell us that.

Leibniz in 1685 had the dream: "when there are disputes among persons, we can simply say: Let us calculate, without further ado, and see who is right"[1]. He was somewhat right. If the dispute is about what is true in the world, indeed one should calculate where the disputes have different predictions and see what is true in the world. Seeing who is right is not just calculation but interaction with the world and observation of the world. If the dispute is over preferences, then no calculation or observation will tell us what is right. If the dispute is about what to do, given preferences and the history of interaction, then his dream is feasible; we should be able to compute who is right.

Here is a decision-theory version of AI utopia. Machines can do science (in what is part of the broad field of machine learning): they interact and observe the world and build theories about how the world works. These theories are open to scrutiny. We don't have preferences over which theory is best: whichever theories best predict the future are the best theories. When politicians want to campaign, they do not campaign on what they will do, they rather campaign on their values. Given the values, and the science of how the world works, computers will then be able to specify what they should do (and even perhaps do it). We have to be scared of a future where there is not this debate about preferences, but we should welcome a future where the debate about values then leads to the best decisions based on these values. Technically we may be able to do this in our lifetime, but socially it will take longer, but it will come.

References

Davis, E. [1990]. *Representations of Commonsense Knowledge*, Morgan Kaufmann, San Mateo, CA.

McCarthy, J. and Hayes, P. J. [1969]. Some philosophical problems from the standpoint of artificial intelligence, *in* M. Meltzer and D. Michie (eds), *Machine Intelligence 4*, Edinburgh University Press, pp. 463–502.

Shanahan, M. [1997]. *Solving the Frame Problem: A Mathematical Investigation of the Common Sense Law of Inertia*, MIT Press, Cambridge, MA.

Sowa, J. F. [2000]. *Knowledge Representation: Logical, Philosophical, and Computational Foundations*, Brooks Cole Publishing Co., Pacific Grove, CA.

Tversky, A. and Kahneman, D. [1974]. Judgment under uncertainty: Heuristics and biases, *Science* **185**: 1124–1131.

[1] http://plato.stanford.edu/entries/leibniz-mind/