Philosophical Agents

John R. Rose • University of South Carolina • rose@cs.sc.edu
Michael N. Huhns • University of South Carolina • huhns@sc.edu

The improvements in Internet-based software agents that are underway at many laboratories and corporations are fulfilling the promise of personalized, friendly Web services. The improvements come at a cost, however—greater implementation complexity. Thus, as we gradually rely more on the improved capabilities of these agents to assist us in networked activities such as e-commerce and information retrieval, we also understand less about how they operate.

Abstraction is the technique we use to deal with complexity. What is the proper kind and level of abstraction for complex software agents? We think it will be reasonable to endow agents with a philosophy. Then, by understanding their philosophies, we can use them more effectively.

For example, consider future NASA missions. As they become longer, more complicated, and farther away, the software systems controlling them will of necessity become larger, more intricate, and increasingly autonomous. Moreover, the missions must succeed in the face of uncertainties, errors, failures, and serendipitous opportunities. While small, well-specified systems with limited types of known external interactions can be proved correct, consistent, and deadlock-free via formal verification, such conditions do not hold for network-based systems. We will basically have to trust the systems, so there should be a principled basis for our trust.

Global System Coherence

An agent-based approach is inherently distributed and autonomous, but when the communication channels that link the agents are noisy or bandwidth-constrained, the agents will have to make decisions locally, which we hope will be coherent globally. We can trust the agents to act autonomously if they embrace ethical principles that we understand and with which we agree.

To endow agents with ethical principles, we as developers need an architecture that supports explicit goals, principles, and capabilities (such as how to negotiate), as well as laws and ways to sanction or punish miscreants. Figure 1 illustrates such an agent architecture that can support both trust and coherence, where coherence is defined as the absence of wasted effort and progress toward chosen goals.

The lowest level of the architecture enables an agent to behave reactively, i.e., react to immediate events. The middle layers are concerned with an agent's interactions with others, while the highest level enables the agent to consider the long-term effects of its behavior on the rest of its society. Agents are typically constructed starting at the bottom of this architecture, with increasingly more abstract reasoning abilities layered on top.

Awareness of other agents and of one's own role in a society, which are implicit at the social commitment level and above, can enable agents to behave coherently. Tambe et al.[1] has shown how a team of agents flying helicopters will continue to function as a coherent team after their leader has crashed, because another agent will assume the leadership role. More precisely, the agents will adjust their individual intentions in order to fulfill the commitments made by the team.

If the agents have sufficient time, they can negotiate about or vote on which agent should become the new leader. When time is short or communication is not allowed, the agents can follow mutually understood social conventions, such as "the agent with the most seniority becomes the new leader."

Ethical Abstractions

Ethics is the branch of philosophy concerned with codes and principles of moral behavior.[2] Most ethical theories distinguish between the concepts of *right* and *good:*

- right is that which is right in itself;
- good is that which is good or valuable for someone or for some end.

 1089-7801/01/$10.00 ©2001 IEEE

The German philosopher, Immanuel Kant (1724–1804), defined the "categorical imperative" as an absolute and universal moral law (of the form "Do this") based entirely on reason (as distinguished from his "hypothetical imperative," which is based on desire: "Do this if you want that"). We can state the categorical imperative in relation to agent behavior as follows: "Agents should act as if the maxim of their action were to secure through their will a universal law of nature." It provides a source of right action. For example, breaking a "promise" is not right, because if all agents did it, the system they support would not function.

Kant's categorical imperative does not contain a way to resolve conflicts of duty. Less stringent formulations specify *prima facie duties*, which do not bind agents absolutely but instead hold generally: "All other things being equal, keep your promises."

So-called *deontological theories*, like Kant's, emphasize "right before good." They oppose the idea that the ends can justify the means, and they place the locus of right and wrong in autonomous adherence to moral laws or duties. These theories distinguish intentional effects from unforeseen consequences. That is, an action is not wrong unless the agent explicitly intends for it to do wrong. This legitimizes inaction, even when inaction has predictable, but unintended effects. For example, consider a NASA deep space probe in which an agent is responsible for managing communications with ground control. The agent would not be wrong to shut down the communications link for diagnostics, even if that severed communications of other agents with ground control.

In contrast, *teleological theories* choose good before right: something is right only if it maximizes the good; in this case, the ends can justify the means. In teleological theories, the correctness of actions is based on how the actions satisfy various goals, not the intrinsic rightness of the actions. Choices of actions can be comparison-based or preference-based.

Egoism is an ethical theory parallel to *utilitarianism*: the utilitarian holds that action should maximize the universal good of all agents; the egoist holds instead that action should maximize self-interest. Both theories are teleological, in that they hold that the right thing to do is always to produce a certain good, where good may be interpreted as various ways:

- pleasure, in which case it is called hedonism;
- preference satisfaction, called micro-economic rationalism (assumes each agent knows its preferences);
- interest satisfaction, called welfare utilitarianism; and
- aesthetic ideals, called ideal utilitarianism.

What agents need to decide actions are not just universal principles (each can be stretched) and not just consequences, but also a regard for their promises and duties. They have prima facie duties to keep promises, help others, repay kindness, and so on.

In the context of a NASA mission, an agent could repay a kindness to another agent by offering, without being asked, to donate some resource such as excess battery power that it has been allocated but does not need. Imagine that the receiving agent already has enough battery power to accomplish its task, but that the additional power gives it an extra safety margin. There is no ranking among these duties, which are highly defeasible. For example, an agent on a NASA deep-space probe might find it acceptable to monopolize a communication channel to ground control to the detriment of other agents because it values the success of its task without regard to the consequences for other agents.

Machined Ethics

Isaac Asimov proposed a moral philosophy for intelligent machines in 1940. His collection of short stories, *I, Robot*,[3] included a *Handbook of Robotics* that defined three Laws of Robotics. These were subsequently augmented in *Foundation and Empire*[4] by the "zeroth law," and the four laws were rewritten as follows:

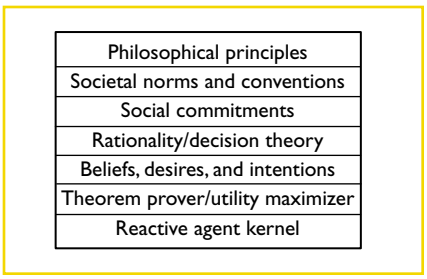

Figure 1. Architecture for a philosophical agent. Layers of deliberation enable socially appropriate agent behavior.

- *Law 0*. A robot may not injure humanity or, through inaction, allow humanity to come to harm.
- *Law 1*. A robot may not injure a human being or, through inaction, allow a human to come to harm.
- *Law 2*. A robot must obey the orders given it by human beings except where such orders would conflict with the First Law.
- *Law 3*. A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.

An adaptation of these laws for a collection of agents sent on a NASA mission might be:

- *Principle 1*. An agent shall not harm the mission through its actions or inactions.
- *Principle 2*. Except where it conflicts with the Principle 1, an agent shall not harm the participants in the mission.
- *Principle 3*. Except where it conflicts with the previous principles, an agent shall not harm itself.
- *Principle 4*. Except where it conflicts with the previous principles, an agent shall make rational progress toward mission goals.
- Principle 5. Except where it conflicts with the previous principles, an agent shall follow established conventions.

- *Principle 6*. Except where it conflicts with the previous principles, an agent shall make rational progress toward its own goals.
- *Principle 7*. Except where it conflicts with the previous principles, an agent shall operate efficiently.

Distributed systems are susceptible to deadlocks and livelocks. However, if the system components obey these seven philosophical principles, then the susceptibilities would disappear, because deadlock and livelock would violate Principle 6.

Nature versus Nurture

What is the source of the robust behavior that a large collection of philosophical agents is expected to show? Wouldn't a collection of agents sharing the same philosophical tenets be susceptible to failing en masse, particularly if they were identical copies? The answer is no, provided the agents are sufficiently intelligent and capable of modifying their reasoning based on their individual perspectives, which in turn are a result of their experiences.

Consider the differences observed between identical twins. Though genetically identical, they develop personality differences as a consequence of their experiences. It is the difference in experience and consequently in perspective that precludes a monoculture of intelligent philosophical agents. On the other hand, if the agents are as dumb as potatoes, then the agent equivalent of an Irish potato famine could well occur.

In a real-world distributed environment, an agent's perception of the world is based on its direct experience and the information it acquires from others. Its perspective is a function of its belief in its own perceptions and what it learns indirectly from the perceptions and beliefs of other agents. Differences in perspective lead to differences in action. Consequently, large collections of intelligent philosophical agents will generate the complex redundancy of diversity and not the simple redundancy of duplication. In this framework, robustness is an emergent property of diversity guided by a common underlying philosophy.

Consider agents modeled on simple ant behavior for path planning.[5] This model, in which robust complex behavior emerges from simple individual agent behavior, does not rely on any hype about future machine intelligence. Instead, there are five simple rules:

- *Rule 1*. Avoid obstacles.
- *Rule 2*. Walk preferentially in the direction of pheromones; otherwise wander randomly.
- *Rule 3*. When carrying food, mark the return path to the next with pheromone (the nest is marked with a distinctive pheromone).
- *Rule 4*. When not carrying food, pick up any food that you find.
- *Rule 5*. When at the nest, drop any food that you are carrying.

The resulting paths are minimum spanning trees that minimize the energy ants expend in gathering food. Almost as an aside, the point is made that ants wandering entirely off on their own will starve or otherwise die.

In our view, there are two critical features to the success of this approach to path planning:

- The number of available ants must be sufficiently large so that ants wandering off in the wrong direction and dying do not threaten overall success.
- The ants must explore different directions looking for food when they do not detect a pheromone path leading to it.

In other words, the robustness of this planning behavior derives from *diverse redundancy* even though the behavior of individual ants is quite simple and they all follow the same principles.

Applying Ethics

A philosophical approach to distributed system design presupposes that the components, or agents, can

- enter into social commitments to collaborate with others,
- change their mind about their results, and
- negotiate with others.

However, the ethical theories we have described are theories of justification, not of deliberation. An agent can decide what basic "value system" to use under any approach.

The deontological theories are narrower and ignore practical considerations, but they are only meant as incomplete constraints—that is, the agent can choose any of the right actions to perform. The teleological theories are broader and include practical considerations, but they leave the agent fewer options for choosing the best available alternative.

All of these ethical approaches are single-agent in orientation and encode other agents implicitly. An explicitly *multiagent* ethics would be an interesting topic for study.

Acknowledgments

This work was supported by the U.S. National Science Foundation under grant no. IIS-0083362.

References

1. M. Tambe, D.V. Pynadath, and N. Chauvat, "Building Dynamic Agent Organizations in Cyberspace," *IEEE Internet Computing*, vol. 4, no. 2, Mar.-Apr. 2000, pp. 65-73.
2. For further reading in applied ethics, see the home page for Carnegie Mellon Center for the Advancement of Applied Ethics at http://www.lcl.cmu.edu/CAAE/index.htm.
3. Isaac Asimov, *I, Robot*, Gnome Books, 1950.
4. Isaac Asimov, *Foundation and Empire*, Gnome Books, 1952.
5. H. Van Dyke Parunak, "'Go to the Ant': Engineering Principles from Natural Multi-Agent Systems," *Annals of Operations Research*, vol. 75, 1997, pp. 69-101.

John R. Rose is a professional cave diver who dabbles in computer science at the University of South Carolina, where his primary research interest is distributed decision networks.

Michael N. Huhns is a professor of computer science and engineering at the University of South Carolina, where he also directs the Center for Information Technology.