Why and How Should Robots Behave Ethically?

Benjamin KUIPERS [1]
Computer Science & Engineering, University of Michigan, USA

Abstract for Plenary Lecture

The prospect of intelligent robotic agents taking increasingly significant roles in our human society suggests that it would be prudent for robot designers to ensure that robot behavior is governed by some sort of morality and ethics — that robots should be *trustworthy*. But what would this actually mean?

Research in artificial intelligence has profited from, and contributed to, the study of human cognition, including the fields of cognitive science and cognitive neuroscience. Likewise, we might hope that efforts to design moral and ethical systems for robots will both draw upon, and contribute to, a deeper understanding of morality, ethics, and trust among human beings [18,11].

Many of the benefits of society come from cooperation, which in turn depends on *trust* between cooperating partners.

> *"Trust is a psychological state comprising the intention to accept vulnerability based upon positive expectations of the intentions or behavior of another."* [15]

For an intelligent robot to function successfully in our society, to cooperate with humans, it must not only be able to act morally and ethically, but it must also be trustworthy. It must earn and keep the trust of humans who interact with it. We will first look at trust among a few partners, who can often choose whether to cooperate, and then at the social norms that make up morality and ethics, that are the foundation of trust and cooperation among everyone in a society.

Describing *decision theory* as the basis for "rational" action, the leading textbook in Artificial Intelligence [16, p.611] says that a rational agent should choose the action that maximizes the agent's expected utility. (*Game theory* is decision theory in contexts where multiple agents are making decisions to maximize their own expected utilities [10].)

The crux of a decision theory or game theory formulation is the definition of the utility function, which is intended to represent the agent's preference over states of the world. In principle, the utility function can be arbitrarily sophisticated, even considering the welfare of everyone in society equally [17]. However, in

[1]University of Michigan, Computer Science & Engineering, 2260 Hayward Street, Ann Arbor, Michigan 48109 USA, Email: kuipers@umich.edu

practice (for simplicity, and following the structure of recreational games), utility is typically defined in terms of the agent's own reward. Unfortunately, examples like the Prisoner's Dilemma [2] and the Tragedy of the Commons [6] demonstrate that "rational" maximization of self-centered expected utility can easily lead to very poor outcomes, both for the individual and for society.

The Prisoner's Dilemma (and even more so, other economic games such as the Public Goods Game or the Basic Trust Game) illustrate barriers to cooperation among a few participants. If every participant contributes their share, everyone gets a good outcome. But each individual participant may do even better by optimizing their own reward at the expense of the others. With self-centered utility functions, each participant "rationally" maximizes their own expected utility, often leading to bad outcomes for everyone.

Naturally, when a potential cooperative partner has a reputation of exploiting any vulnerability to maximize selfish reward, trust toward that partner is not well justified, and the benefits of cooperation are not likely to ensue. We will review methods that have been proposed for incorporating trust and trustworthiness into utility functions to reward trustworthiness in game theory.

We turn our attention to morality and ethics, which define social norms that provide a basis for trust among all the members of society. As more people follow the social norm, "*Thou shalt not kill*", each of us needs to spend less effort and vigilance on defending ourselves, leaving more resources for more productive uses [14]. Likewise for social norms like "*Thou shalt not steal*", or even "*Drive on the right side of the road*".

I will start from the position that the purpose of morality and ethics is to improve the welfare of society as a whole. A society is made up of individual agents, who have some degree of autonomy, some degree of interdependence, and some ability to choose whether to participate in the society. Therefore, the welfare of the individual agents is also important, but is, in an evolutionary sense, secondary to the welfare of the society as a whole.

This is a consequentialist position on the fundamental nature of morality and ethics. From this perspective, the ultimate determination of whether the welfare of society has been well served is whether the society survives and thrives, including whether it is able to generate viable successor societies as conditions change [4,13]. This evolutionary perspective suggests that what is right or wrong, or good or bad, may change with conditions.

Of course, it is completely infeasible for an individual agent to make accurate predictive consequentialist calculations of the effects of potential actions on the long-term welfare of society as a whole. Therefore, individual agents need feasible methods for making individual moral decisions in real time. These are necessarily *heuristics* — methods that provide useful, practical approximations of the intractable ideal criteria.

As we have discussed, one type of heuristic method that is particularly favored in artificial intelligence includes *decision theory* and *game theory*, which define the "rational" choice of action as the one that maximizes the agent's expected utility [16]. We treat decision theory and game theory as heuristic methods in this context because, to be tractable, they must be applied to models that are vastly

simpler than the complex reality that they attempt to describe. These heuristic methods are clearly closely related to *utilitarianism*.

Another type of heuristic method consists of pattern-directed rules and constraints that can evaluate a situation quickly enough to support real-time decision-making. Constraints that rule out certain actions as morally unacceptable (e.g., testifying against your partner in crime in return for a reduced sentence), transform a game like the Prisoner's Dilemma into a different game with an obvious optimal solution. Rules can lead to preferences over potentially acceptable actions by classifying them as positive or negative along various dimensions of moral evaluation [5]. This heuristic resembles the ethical theory of *deontology*.

A third type of heuristic method, *case-based reasoning* [8], uses knowledge in the form of concrete cases that illustrate positive and negative examples of particular natural concepts. To discern which actions best exemplify virtuous behavior in the current situation, a case-based reasoning method assesses the similarities between proposed actions and the scenarios exemplified by the cases. The action retrieved from the best-matching case is then modified appropriately for the current situation. This heuristic is closely related to the philosophical method of *casuistry* [7] and to *virtue ethics* [1].

One theme in artificial intelligence research is that effective commonsense knowledge may be structured as multiple distinct representations for the same domain [12]. My own work on spatial knowledge and the Spatial Semantic Hierarchy [9,3] shows how a robot can use topological, metrical, procedural, and continuous control methods, each when most appropriate, to do different kinds of learning, planning, and acting in the spatial environment.

In the moral domain, these three very different heuristics can be used to examine proposed solutions to a moral problem, ruling out solutions that are unacceptable according to one or another criterion, and providing a useful ordering on the remaining viable options.

- *Should you use a sharp knife to cut into the body of a human being?* Of course not, unless you are a qualified surgeon performing a necessary operation. (Deontology: a rule with an exception.)
- *If you are that surgeon, is it permissible to sacrifice this patient in order to save the lives of five others?* Of course not! (Virtue ethics: a good surgeon keeps faith with the patient.)
- *Is it OK to throw the switch that saves five lives by directing a runaway trolley onto a side track, where it will kill one person who would have been safe?* Well, ... (Deontology says it's wrong to allow preventable deaths; Utilitarianism says fewer deaths is better; Virtue ethics says the virtuous person can make hard choices.)

I argue that heuristics based on utilitarianism (decision theory), deontology (rule-based and constraint-based systems), and virtue ethics (case-based reasoning) are all important tools in the toolkit for creating artificial agents capable of participating successfully in our society. Each tool is useful in certain contexts, and perhaps less useful in others.

References

[1] Julia Annas. *Intelligent Virtue*. Oxford University Press, 2011.

[2] Robert Axelrod. *The Evolution of Cooperation*. Basic Books, 1984.

[3] P. Beeson, J. Modayil, and B. Kuipers. Factoring the mapping problem: Mobile robot map-building in the Hybrid Spatial Semantic Hierarchy. *International Journal of Robotics Research*, 29(4):428–459, 2010.

[4] Daniel C. Dennett. *Darwin's Dangerous Idea*. Simon & Schuster, 1995.

[5] Jonathan Haidt. *The Righteous Mind: Why Good People are Divided by Politics and Religion*. Vintage Books, NY, 2012.

[6] Garrett Hardin. The tragedy of the commons. *Science*, 162:1243–1248, 1968.

[7] Albert R. Jonson and Stephen Toulmin. *The Abuse of Casuistry: A History of Moral Reasoning*. University of California Press, 1988.

[8] Janet Kolodner. *Case-Based Reasoning*. Morgan Kaufmann, 1993.

[9] B. Kuipers. The Spatial Semantic Hierarchy. *Artificial Intelligence*, 119:191–233, 2000.

[10] Kevin Leyton-Brown and Yoav Shoham. *Essentials of Game Theory*. Morgan & Claypool, 2008.

[11] Patrick Lin, Keith Abney, and George A. Bekey, editors. *Robot Ethics: The Ethical and Social Implications of Robotics*. MIT Press, 2012.

[12] Marvin Minsky. *The Society of Mind*. Simon & Schuster, NY, 1985.

[13] Ara Norenzayan. *Big Gods: How Religion Transformed Cooperation and Conflict*. Princeton University Press, 2013.

[14] Steven Pinker. *The Better Angels of Our Nature: Why Violence Has Declined*. Viking Adult, 2011.

[15] D. M. Rousseau, S. B. Sitkin, R. S. Burt, and C. Camerer. Not so different after all: a cross-discipline view of trust. *Academy of Management Review*, 23(3):393–404, 1998.

[16] S. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach*. Prentice Hall, third edition, 2010.

[17] Peter Singer. *The Expanding Circle: Ethics, Evolution, and Moral Progress*. Princeton University Press, 1981.

[18] Wendell Wallach and Colin Allen. *Moral Machines: Teaching Robots Right from Wrong*. Oxford University Press, 2009.