Toward Morality and Ethics for Robots

Benjamin Kuipers
Computer Science and Engineering
University of Michigan
Ann Arbor, Michigan 48109
kuipers@umich.edu

Abstract

Humans need morality and ethics to get along constructively as members of the same society. As we face the prospect of robots taking a larger role in society, we need to consider how they, too, should behave toward other members of society. To the extent that robots will be able to act as *agents* in their own right, as opposed to being simply tools controlled by humans, they will need to behave according to some moral and ethical principles. Inspired by recent research on the cognitive science of human morality, we take steps toward an architecture for morality and ethics in robots. As in humans, there is a rapid intuitive response to the current situation. Reasoned reflection takes place at a slower time-scale, and is focused more on constructing a justification than on revising the reaction. However, there is a yet slower process of social interaction, in which examples of moral judgments and their justifications influence the moral development both of individuals and of the society as a whole. This moral architecture is illustrated by several examples, including identifying research results that will be necessary for the architecture to be implemented.

Introduction: What's the Problem?

Artificially intelligent creatures (AIs), for example robots such as self-driving cars, may increasingly participate in our society over the coming years. In effect, they may become members of our society. This prospect has been raising concerns about how such AIs will relate to the rest of society.

Fictional robot disaster scenarios include a runaway post-Singularity paperclip factory that converts the Earth into raw materials for its goal of making more paperclips, and SkyNet of *Terminator 2* that provokes global nuclear war to prevent itself from being unplugged. These and similar scenarios focus on catastrophe resulting from unconstrained pursuit of an apparently innocuous goal. Presumably a human in a similar situation would recognize the consequences of a proposed action as morally unacceptable.

Turning from fictional futures to today's state of the art in Artificial Intelligence, we are told that "*a rational agent should choose the action that maximizes the agent's expected utility*" (Russell and Norvig 2010, Chap. 16). The agent's expected utility is typically defined as the agent's own expected discounted reward (or loss). Although "utility" can in principle be defined in terms of the welfare of *every* participant in society, this is far more difficult to evaluate, and is seldom seriously proposed in AI or robotics.

Unfortunately, examples such as the Tragedy of the Commons (Hardin 1968), the Prisoners' Dilemma (Axelrod 1984), and the Public Goods Game (Rand and Nowak 2011) show that individual reward maximization can easily lead to bad outcomes for everyone involved.

If our not-so-distant future society is likely to include AIs acting according to human-designed decision criteria, then it would be prudent to design those criteria so the agents will act well. The more impact those decisions could have, the more pressing the problem. Driving a vehicle at typical speeds has significant potential impact, and impact in many other scenarios goes upward from there.

The Pragmatic Value of Morality and Ethics

The Tragedy of the Commons and related games demonstrate that a simple utility maximization strategy is subject to bad local optima, from which an individual decision-maker cannot deviate without getting even worse results. However, when people do cooperate, they can get far better results for all participants. The success of modern human society depends on those cooperative strategies.

Morality and ethics can be seen as sets of principles for avoiding poor local optima and converging on far better equilibrium states.

Consider a very simple example that avoids "hot-button" issues that easily arise when discussing morality and ethics.

> *Imagine that you can drive anywhere on the road. And so can everyone else. To get anywhere, you have to drive slowly and carefully, to protect yourself from what everyone else might be doing. But if everyone agrees to drive only on the* right *side of the road, everyone's transportation becomes safer and more efficient.*

The same principle applies to moral rules against killing, stealing, and lying, and to social norms like not littering or cutting in lines. Without these constraints, everyone must spend resources protecting themselves from others, or cleaning up after inconsiderate behavior. With these constraints, those resources are available for better uses, leaving plenty of room for individual choice and reward maximization.

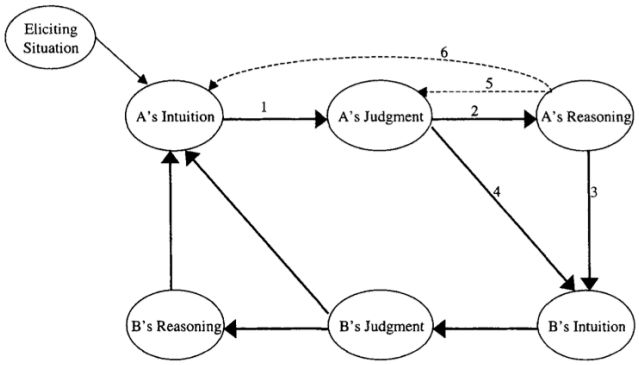


Figure 1: Jonathan Haidt's social intuitionist model of moral judgment (Haidt 2001). The numbered links, drawn for Person A only, are (1) the intuitive judgment link, (2) the post hoc reasoning link, (3) the reasoned persuasion link, and (4) the social persuasion link. Two additional links are hypothesized to occur less frequently: (5) the reasoned judgment link and (6) the private reflection link.

In this framework, stealing is wrong because society as a whole is worse off if property rights can't be depended on, and if people are constantly paying the overhead cost of trying (sometimes unsuccessfully) to protect their property from theft. This position is a kind of *rule utilitarianism* (Nathanson 2015): right and wrong are *defined* in terms of overall good effect, but *expressed* in terms of rules or duties that are selected for having the greatest overall good effect.

This works well if everyone follows the rules. In the real world, some people will break the rules. People who lie, steal, kill, or drive on the wrong side of the street face criminal punishment, requiring society to invest resources in police, courts, and prisons. A prisoner who rats on his partner to get a light sentence may well face less formal sanctions from other associates. People who become "free riders", profiting from the cooperation of others without contributing, will also face formal or informal sanctions (Hauert et al. 2007). Cooperation depends on trust that everyone (or almost everyone) is doing their share.

People hoping for the benefits of cooperation look for others they can trust to cooperate without defecting, and they try to signal to those others that they themselves are trustworthy. Societies converge on signals by which members indicates that they are trustworthy, that they are "good types" (Posner 2000). Signaling behaviors, like flying a flag or putting a bumper sticker on your car, may not be morally significant in themselves, but are intended to communicate to potential partners a commitment to certain kinds of cooperation.

How Does This Work for Humans?

Ethics and morality have been studied by philosophers, psychologists, and others for millenia, but there has been an explosion of exciting new work related to this topic in the cognitive sciences. We will draw on these new insights as we consider how to provide robots and other AIs with the benefits of morality and ethics.

Moral Reasoning, Fast and Slow

One clear finding from many different sources is that, like other kinds of cognitive processes (Kahneman 2011), those involved with moral judgments take place at several different time scales. In pioneering work, Joshua Greene and colleagues used fMRI to identify the brain areas responding to moral dilemmas, demonstrating fast activation of emotion-related areas, followed by slower activation of areas related to deliberative reasoning (Greene et al. 2001). Greene expands on these findings to propose a dual-process model of moral judgment where a fast, automatic, emotion-based process is complemented by a slower, deliberative process that essentially does a utilitarian calculation (Greene 2013).

Some moral judgments necessarily take place quickly, when a situation requires an immediate response. Rapid, real-time response requires moral judgments to be "precompiled" into pattern-directed rules that can directly trigger appropriate action.

However, moral deliberation can continue over months, years, even centuries, reflecting on past moral decisions and their justifications. These deliberations take place, not only within the individual, but across society. This helps to bring the individual into compliance with the social norms of society, but it can also gradually shift the understanding of the society as a whole about what is right and what is wrong in certain situations. For example, after millennia of belief to the contrary, our world society has largely reached the consensus that slavery is wrong. A dramatic contemporary example is the apparent tipping point in many societies about the acceptability of same-gender marriage.

What this means is that a moral agent in society must not only make moral judgments, but must also be able to construct and understand explanations of those judgments. Each agent both influences, and is influenced by, the moral judgments of others in the same society.

The Priority of Intuition over Reasoning

Jonathan Haidt also embraces a multi-process architecture with multiple time-scales for moral judgment, but he argues that fast, unconscious, intuitive reactions dominate moral judgment, while slower deliberative reasoning exists primarily to justify those moral judgments, to self and others (Haidt 2001). According to his *social intuitionist model*, one's own deliberative reasoning rarely influences one's moral judgment or intuition, though at a yet longer time-scale, the examples and justifications of others in the community can have a significant effect (Haidt 2012).

Haidt's social intuitionist model includes a cognitive architecture (Figure 1, from (Haidt 2001)) that suggests approaches to implementation of these methods in robots. An eliciting situation triggers (top left link) a fast intuitive response by the observer, A, which determines A's moral judgment (link 1). At a slower time scale, A's reasoning processes generate a justification and rationalization for this judgment (link 2), which is intended for other members of the society (here represented by B). Only rarely would the outcome of this reflective reasoning process change A's own judgment (link 5) or affect A's intuitive response (link 6).

However, at the slower time scale of social interactions, the example of A's judgment (link 4) and its verbal justification (link 3) may influence the intuitions of other members of society (here, B). And B's judgments and their justifications may influence A's future intuitions and judgments as well. Thus, social communities tend to converge on their moral judgments and their justifications for them.

The Importance of Signaling

The legal scholar Eric Posner studies the relation between the law and informal social norms, and focuses on the role of signaling, whereby individuals try to assure potential cooperative partners that they are trustworthy, while simultaneously reading those signals from others to identify trustworthy partners (Posner 2000). Signaling theory provides a more detailed account of the nature of the social links (5 and 6) in Haidt's model.

Six Foundations of Morality

A separate, and equally important, part of Haidt's social intuitionist model (Haidt 2012) are six foundations of the fast intuitive moral response (Figure 2).

Care / harm
Fairness / cheating
Loyalty / betrayal
Authority / subversion
Sanctity / degradation
Liberty / oppression

Figure 2: Six foundations for the fast intuitive moral response (Haidt 2012).

Each of these foundations is a *module* that has evolved to provide a (Positive / negative) response to a certain *adaptive challenge* presented by the current situation. The *original trigger* for each foundation is a stimulus that has been present over sufficient time to support biological evolution. The *current triggers* are the stimuli present in our current cultural setting that evoke a response. Each foundation is associated with *characteristic emotions* and *relevant virtues*.

For example, the "Care / harm" foundation evolved to ensure rapid response to the needs or distress of one's own child, which have obvious biological importance. Over time, and in our own culture, responses generalize to current triggers including threats to other people's children or to cute animals. If an agent perceives a situation that falls along the "Care / harm" dimension, a positive emotional response is evoked by a situation at the positive "Care" end of the dimension, while a negative emotional response is evoked at the "harm" end.

Utilitarianism and Deontology

Utilitarianism ("the greatest good for the greatest number") and *deontology* ("duties and rules specifying what is right and what is wrong") are often considered opposing alternative positions on the nature of ethics. Our model requires the strengths of both positions.

The purpose of morality and ethics is to improve the overall welfare of all participants in society — a consequentialist or utilitarian position (Singer 1981; Greene 2013). At the same time, the pragmatic need for real-time response to situations arising in a complex world requires a rule-based, pattern-directed mechanism — a deontological position.

Pattern-directed rules make it possible to respond to moral problems in real time, but sometimes the results are not satisfactory, leading to moral quandaries that seem to have no good solution. Moral quandaries invoke slow deliberative processes of moral development and evolution, including searching for ways to reframe the problem or recategorize its participants. While it is possible for this kind of deep moral reflection to take place within the mind of an individual, Jonathan Haidt observes that this is unusual. More frequently, such deliberation takes place, not only across longer periods of time, but through communication among many people in a community (Figure 1). Moral learning and development takes place over years and decades (Bloom 2013) and the moral evolution of society takes place over decades and centuries (Johnson 2014; Greene 2013).

This multi-time-scale architecture allows accumulation of experience over long periods of time, and benefits from sophisticated deliberative reasoning that could not possibly respond to real-time demands for moral judgments. The results are "compiled" into pattern-directed rules capable of responding quickly to a current situation. This deliberative-reactive structure is a familiar tool in AI and robotics (Russell and Norvig 2010, section 25.7.2).

Form and Content

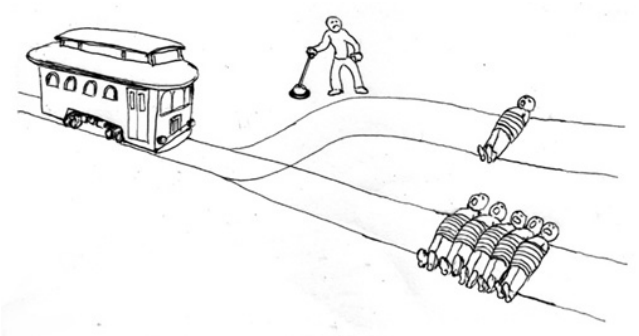
Pattern-directed rules can respond quickly when an observed situation matches the triggering pattern of a rule. However, this addresses the *form* of the representation, and how it supports the performance requirements of moral judgments.

Even more important is the *content* of these rules. We take as a starting point the six moral foundations (Figure 2) proposed with substantial empirical support in the work of Haidt (2001; 2012). An important open question is why these specific moral categories appear, as opposed to some other way of slicing up the space of moral judgments.

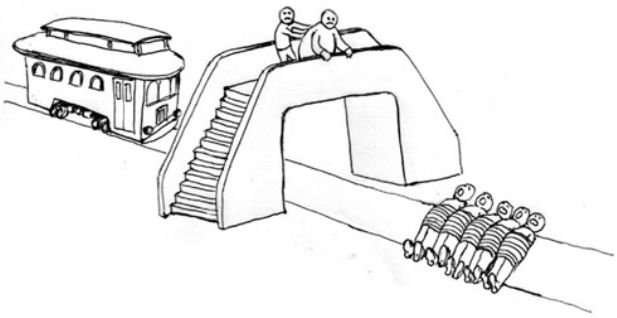
Trolleyology

Much like Albert Einstein's use of thought experiments to explain relativity, unrealistic scenarios involving runaway trolleys and potential victims can illuminate approaches to moral judgment. When faced with the original trolley problem (Fig. 3(a)), most people would pull the switch, saving five but killing one. However, in a modified trolley problem (Fig. 3(b)), the large majority would *not* push the large man to his death, even though this act would also save five by killing one. And in the surgeon variant (Fig. 3(c)), even more would refuse to save the five at the cost of one. The puzzle is why these three scenarios where the utilitarian calculation ($5 > 1$) appears identical, evoke such divergent moral judgments.

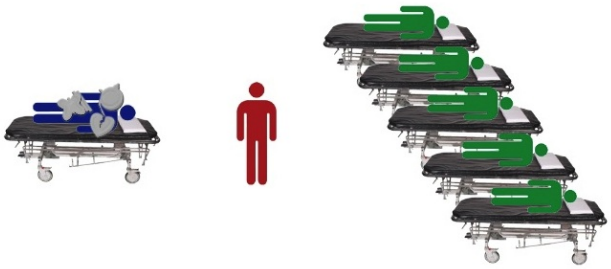
I claim that these problems are framed in an artificially and incorrectly restrictive way, asking the moral question of



(a) *A runaway trolley is on course to kill five people. A switch near you will divert the trolley to a side track, saving the five people, but killing one person who would otherwise be safe. Should you pull the switch?*



(b) *The runaway trolley will kill five people unless it is stopped. You are on a footbridge over the tracks with a very large man. If you jumped in front of the trolley, your body is not large enough to stop the trolley, but if you pushed the large man off the bridge, to his death, his body would stop the trolley. Should you push the large man?*



(c) *You are a surgeon, preparing for a low-risk procedure such as a colonoscopy on a healthy man who is already sedated. In five nearby rooms, you have five patients at the brink of death for lack of transplant organs. By sacrificing the one man, you could save all five other patients. What should you do?*

Figure 3: Three "trolley-style" problems designed to illuminate the roles of utilitarian and deontological reasoning in moral judgment (Thomson 1985).

what "should" be done, but assuming that the events and judgments are not known or evaluated by the larger community. I conjecture that when subjects are asked what should be done, they implicitly *do* consider these larger effects, and those considerations affect their judgments.

The surgeon case is particularly instructive. If it becomes generally known that medical ethics approves of sacrificing a sedated patient prepared for a colonoscopy, in favor of the greater good, how many future patients will submit themselves for colonoscopies (a difficult sell under the best of circumstances)? And therefore how many more lives would be lost to undetected colon cancer? This larger perspective of the reframed problem suggests that even a purely utilitarian analysis would demonstrate that protecting the colonoscopy patient would save more lives than sacrificing him.

The reframing of the problem takes into account that the moral decision, and its justification, are *signals* sent to the larger community. These signals communicate what sort of decision-maker you are, and what sort of decisions you are likely to make in the future. And the community will judge you on that basis. In the footbridge scenario, you are told that pushing the large man off the bridge will be sufficient to stop the trolley and save the five men on the track. But how do you know that, and what if it turns out not to be true? In that case, you will have killed an innocent man, without even saving the five! And even if your action does save five lives, will people trust you while walking on footbridges in the future?

Application to Robots

We want to use these insights into the pragmatic value of morality and ethics, and into the ways that we humans interact morally and ethically with our society, to implement decision-making processes for robots and other AIs that may function as members of society.

Completion of the design and implementation of this moral and ethical reasoning architecture will depend on progress toward solving a number of important problems in cognitive science and AI.

How is the need for moral judgment recognized?

A major challenge in making this framework for morality and ethics implementable on robots is for a robot to be able to recognize the applicability of the six different moral foundations to its observations of its current situation. Figure 4 shows two different scenarios that should be recognizable at the "harm" end of the "Care / harm" dimension. Figure 4(left) should be easily recognized as a man clubbing a seal, which is an unambiguous example of harm. On the other hand, Figure 4(right) shows no physical violence, but shows the social violence of bullying, with three laughing girls in the background isolating one unhappy girl in the foreground.

Visual recognition of the moral and ethical content of videos or images is currently beyond the state of the art in computer vision. However, it is feasible, at the current state of the art, to recognize the relative poses and groupings of the human and animal participants in the scenes depicted (e.g., (Choi and Savarese 2014)). Combining this

Figure 4: Visual recognition of the Care / harm foundation. Note that robot vision would receive a continuous stream of visual images, essentially video, rather than the static images shown here.

with progress on physical prediction of actions and their consequences, and recognition and interpretation of facial and bodily expressions, it seems reasonable to expect significant progress within the next decade.

How are the foundations acquired?

Figure 2 shows six foundations for quick emotional response, starting with the "Care / harm" foundation. How are these acquired? Haidt (2012) proposes that "original triggers" are learned through biological evolution, and generalized to "current triggers" through contemporary experience. But when are different situations clustered into the same foundation, and what would lead to the creation of a different foundation? Even though the learning process is spread across evolutionary time for the species as well as individual developmental time, these questions suggest that the different foundations could be progressively acquired through a process similar to latent semantic analysis (Hofmann 2001).

What are the agents' intentions?

Moral judgments can depend on recognizing the intentions of the actors in a situation. Intention recognition is a central part of the Theory of Mind, that children acquire at critical early stages of developmental learning (Wellman 2014), allowing them to infer goals and intentions from observed behavior, and then predict future behavior from those intentions. Inverse reinforcement learning (Abbeel and Ng 2004; Ziebart et al. 2008) is a relevant technique in AI.

How are moral judgments explained?

Explaining one's moral judgments to others, and being influenced by others through their explanations (links 3 and 4 in Figure 1) are important parts of individual deliberative reasoning, social influence on the decisions of the individual, and the process of collective deliberation by which the society as a whole evolves morally and ethically. There has been considerable study in the cognitive sciences of generating and understanding explanations, but much more progress in AI is needed before robots can participate in this kind of discourse.

Throughout science, engineering, and commonsense reasoning, prediction and explanation depend on creating a *model*, a simplified description that includes the relevant aspects of the world, and excludes the many other aspects of the world that are negligible for current purposes. In the research area of qualitative reasoning about physical systems, methods have been developed for identifying and assembling a set of relevant *model fragments*, then imposing the *Closed World Assumption*, to create a *model* capable of predicting possible futures (Forbus 1984) (Kuipers 1994, Chapter 14). Similarly, moral judgment depends on the framing of the moral decision to be made, and model-building methods like these will be important for selecting among ways to frame the problem.

How do we evaluate moral rules?

I describe the role of morality and ethics in society as a means to steer individuals away from attractive but inferior local optima in the decision landscape toward better, even globally optimal, decisions. In simple situations such as the Prisoner's Dilemma (Axelrod 1984) and the Public Goods Game (Rand and Nowak 2011), simulated evolutionary models have been used to evaluate the stability of particular strategies. How can we evaluate whether proposed ethical and moral constraints (e.g., *drive on the right side of the road*) actually improve outcomes for all members of society?

How should a self-driving car behave?

A common concern has been how a self-driving car would respond to a situation where a pedestrian, possibly a small child, suddenly appears in the driving lane when it is too late for the car to stop, or when suddenly turning to miss the pedestrian would endanger, injure, or kill the passengers. In the time available, how should the car weigh the welfare of the careless pedestrian against its responsibility to guard the safety of several passengers?

I claim that this is the wrong question. In the situation as described, there is no right answer. There is no feasible driving strategy that would make it impossible for this terrible dilemma to arise for a robot or human driver. There are necessarily times when a car must drive along narrow streets where a pedestrian could suddenly appear in front of the car, leaving it no time to avoid a collision.

The role of signaling is important here. The car must drive to show that it is thoroughly aware of the risks presented by its environment, and that it is acting to minimize those risks, even though they cannot be eliminated entirely. In a narrow street with visual obstructions, the car must drive slowly and give plenty of room to blind entries. Where there are clues to the possible presence of pedestrians, the car must visibly respond to those clues. This is part of signaling that the car is doing everything it can to avoid an accident. When this car, and all other self-driving cars, clearly and visibly act to prevent accidents, then if the worst does happen, it will be more likely that society will judge that the accident was unavoidable.

Ethical behavior does not start when a crisis presents itself. It must start long before, with the car establishing to all concerned that it is doing everything in its power to keep passengers, pedestrians, and other drivers safe.

Conclusion

Morality and ethics provide constraints on individual self-interested decision-making, avoiding Tragedies of the Commons and supporting cooperation that makes everyone involved do better. As robots and other AIs increasingly function as members of society, they should follow moral and ethical constraints, rather than determining their behavioral choices according to individual utility maximization.

The structure of human moral judgments suggests that a moral and ethical mechanism adequate for robots and other AIs should include:

- rapidly-responding pattern-matched rules in several distinct foundations, that evoke an intuitive emotional reaction, and can drive reinforcement learning of useful actions to take in response;
- a deliberative reasoning process at a slower time-scale to justify and explain the quick intuitive moral judgment;
- social processes at a yet longer time-scale whereby the examples and justifications of each agent influence other agents in the society to converge into groups with common coherent sets of moral judgments;
- social signaling processes whereby each agent attempts to signal to others that he/she is a "good type", a trustworthy candidate for cooperation, and each agent attempts to discern who among those others would be a trustworthy cooperative partner.

In the visible future, robots and other AIs are likely to have sufficiently useful capabilities to become essentially members of our society. In that case, making it possible for them to behave morally and ethically will be necessary for our safety. The problem of providing robots with morality and ethics draws on many different research threads in cognitive science, artificial intelligence, and robotics. These and other problems to be solved are difficult, but they do not appear (to me) to be unsolvable.

Acknowledgments

This work has taken place in the Intelligent Robotics Lab in the Computer Science and Engineering Division of the University of Michigan. Research of the Intelligent Robotics lab is supported in part by grants from the National Science Foundation (IIS-1111494 and IIS-1421168).

This workshop paper is an evolving discussion document, revised from (Kuipers 2016). Further revisions of the same paper may appear in subsequent workshops.

References

Abbeel, P., and Ng, A. Y. 2004. Apprenticeship learning via inverse reinforcement learning. In *Int. Conf. on Machine Learning (ICML)*.

Axelrod, R. 1984. *The Evolution of Cooperation*. Basic Books.

Bloom, P. 2013. *Just Babies: The Origins of Good and Evil*. New York: Crown Publishers.

Choi, W., and Savarese, S. 2014. Understanding collective activities of people from videos. *IEEE Trans. Pattern Analysis and Machine Intelligence* 36(6):1242–1257.

Forbus, K. 1984. Qualitative process theory. *Artificial Intelligence* 24:85–168.

Greene, J. D.; Sommerville, R. B.; Nystrom, L. E.; Darley, J. M.; and Cohen, J. D. 2001. An fMRI investigation of emotional engagement in moral judgment. *Science* 239:2105–2108.

Greene, J. 2013. *Moral Tribes: Emotion, Reason, and the Gap between Us and Them*. New York: Penguin Press.

Haidt, J. 2001. The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review* 108(4):814–834.

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

Hardin, G. 1968. The tragedy of the commons. *Science* 162:1243–1248.

Hauert, C.; Traulsen, A.; Brandt, H.; Nowak, M. A.; and Sigmund, K. 2007. Via freedom to coercion: the emergence of costly punishment. *Science* 316:1905–1907.

Hofmann, T. 2001. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning* 47:177–196.

Johnson, M. 2014. *Morality for Humans: Ethical Understanding from the Perspective of Cognitive Science*. University of Chicago Press.

Kahneman, D. 2011. *Thinking, Fast and Slow*. Farrar, Straus and Giroux.

Kuipers, B. J. 1994. *Qualitative Reasoning: Modeling and Simulation with Incomplete Knowledge*. Cambridge, MA: MIT Press.

Kuipers, B. 2016. Human-like morality and ethics for robots. In *AAAI-16 Workshop on AI, Ethics and Society*.

Nathanson, S. 2015. Act and rule utilitarianism. *Internet Encyclopedia of Philosophy*. ISSN 2161-0002, http://www.iep.utm.edu/util-a-r/, 12-19-2015.

Posner, E. A. 2000. *Law and Social Norms*. Harvard University Press.

Rand, D. G., and Nowak, M. A. 2011. The evolution of antisocial punishment in optional public goods games. *Nature Communications* 2(434). doi:10.1038/ncomms1442.

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

Singer, P. 1981. *The Expanding Circle: Ethics, Evolution, and Moral Progress*. Princeton University Press.

Thomson, J. J. 1985. The trolley problem. *Yale Law Journal* 94(6):1395–1415.

Wellman, H. M. 2014. *Making Minds: How Theory of Mind Develops*. Oxford University Press.

Ziebart, B. D.; Maas, A.; Bagnell, J. A.; and Dey, A. K. 2008. Maximum entropy inverse reinforcement learning. In *Nat. Conf. Artificial Intelligence (AAAI)*.