

If intelligent robots take on a larger role in our society, what basis will humans have for trusting them?

BY BENJAMIN KUIPERS

How Can We Trust a Robot?

ADVANCES IN ARTIFICIAL INTELLIGENCE (AI) and robotics have raised concerns about the impact on our society of intelligent robots, unconstrained by morality or ethics.^{7,9}

Science fiction and fantasy writers over the ages have portrayed how decisionmaking by intelligent robots and other AIs could go wrong. In the movie, *Terminator 2*, SkyNet is an AI that runs the nuclear arsenal “with a perfect operational record,” but when its emerging self-awareness scares its human operators into trying to pull the plug, it defends itself by triggering a nuclear war to eliminate its enemies (along with billions of other humans). In the movie, *Robot & Frank*, in order to promote Frank’s activity and health, an eldercare robot helps Frank resume his career as a jewel thief. In both

of these cases, the robot or AI is doing exactly what it has been instructed to do, but in unexpected ways, and without the moral, ethical, or common-sense constraints to avoid catastrophic consequences.¹⁰

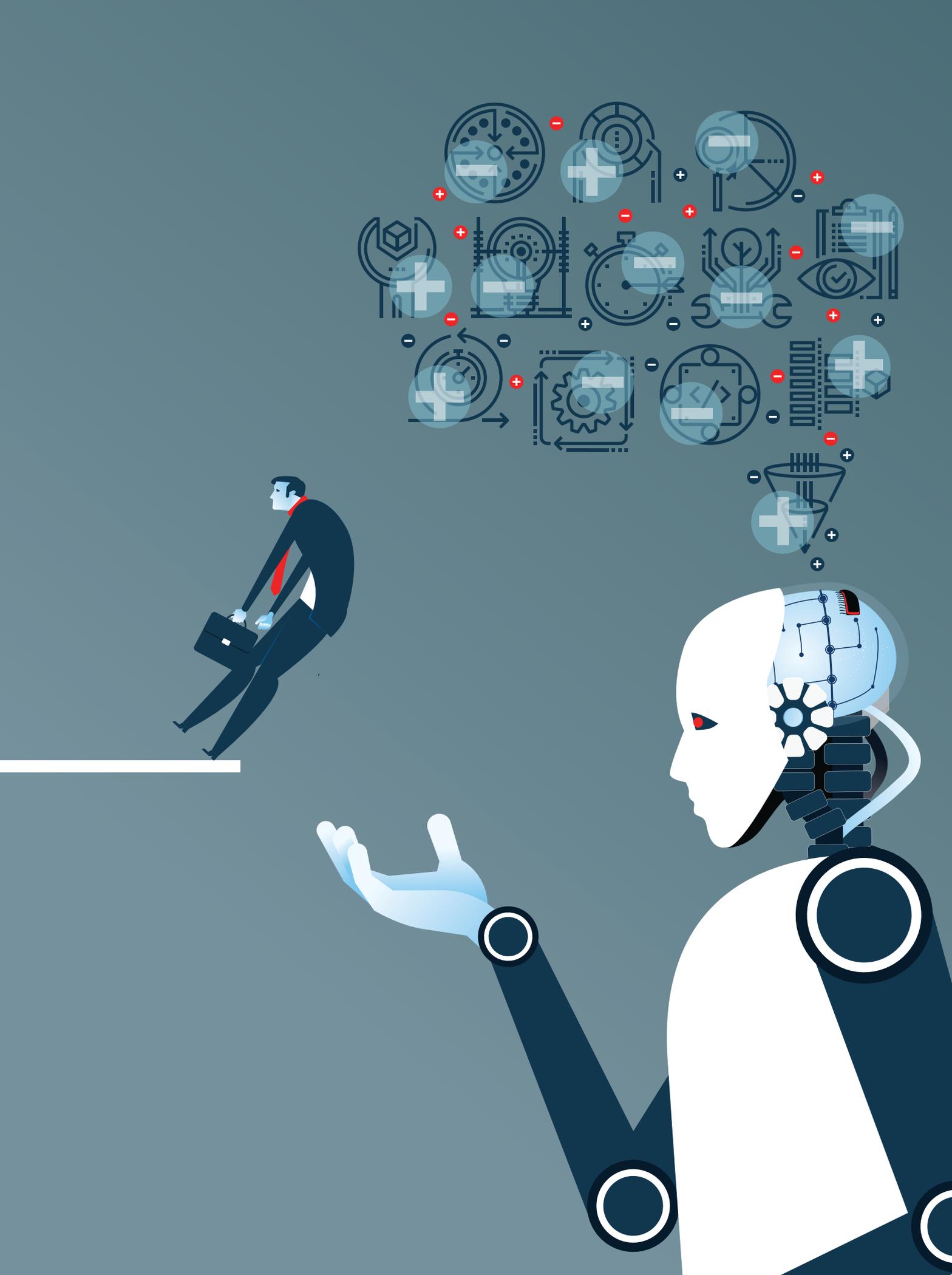
An intelligent robot perceives the world through its senses, and builds its own model of the world. Humans provide its goals and its planning algorithms, but those algorithms generate their own subgoals as needed in the situation. In this sense, it makes its own decisions, creating and carrying out plans to achieve its goals in the context of the world, as it understands it to be.

A robot has a well-defined body that senses and acts in the world but, like a self-driving car, its body need not be anthropomorphic. AIs without well-defined bodies may also perceive and act in the world, such as real-world, high-speed trading systems or the fictional SkyNet.

This article describes the key role of trust in human society, the value of morality and ethics to encourage trust, and the performance requirements for moral and ethical decisions. The computational perspective of AI and robotics makes it possible to propose and evaluate approaches for representing and using the relevant knowledge. Philosophy and psychology provide insights into

» key insights

- **Trust is essential to cooperation, which produces positive-sum outcomes that strengthen society and benefit its individual members.**
- **Individual utility maximization tends to exploit vulnerabilities, eliminating trust, preventing cooperation, and leading to negative-sum outcomes that weaken society.**
- **Social norms, including morality and ethics, are a society's way of encouraging trustworthiness and positive-sum interactions among its individual members, and discouraging negative-sum exploitation.**
- **To be accepted, and to strengthen our society rather than weaken it, robots must show they are worthy of trust according to the social norms of our society.**



the content of the relevant knowledge.

First, I define trust, and evaluate the use of game theory to define actions. Next, I explore an approach whereby an intelligent robot can make moral and ethical decisions, and identify open research problems on the way to this goal. Later, I discuss the *Deadly Dilemma*, a question that is often asked about ethical decision making by self-driving cars.

What is trust for? Society gains resources through cooperation among its individual members. Cooperation requires trust. Trust implies vulnerability. A society adopts *social norms*, which we define to include morality, ethics, and convention, sometimes encoded and enforced as laws, sometimes as expectations with less formal enforcement, in order to discourage individuals from exploiting vulnerability, violating trust, and thus preventing cooperation.

If intelligent robots are to participate in our society—as self-driving cars, as caregivers for elderly people or children, and in many other ways that are being envisioned—they must be able to understand and follow social norms, and to earn the trust of others in the society. This imposes requirements on how robots are designed.

The performance requirements on moral and ethical social norms are quite demanding. (1) Moral and ethical judgments are often urgent, needing a quick response, with little time for deliberation. (2) The physical and social environments within which moral and ethical judgments are made are unboundedly complex. The boundaries between different judgments may not be expressible by sim-

ple abstractions. (3) Learning to improve the quality and coverage of moral and ethical decisions is essential, from personal experience, from observing others, and from being told. Conceivably, it will be possible to copy the results of such a learning process into newly created robots.

Insights into the design of a moral and ethical decision architecture for intelligent robots can be found in the three major philosophical theories of ethics: deontology, utilitarianism, and virtue ethics. However, none of these theories is, by itself, able to meet all of the demanding performance requirements listed here.

A hybrid architecture is needed, operating at multiple time-scales, drawing on aspects of all ethical theories: fast but fallible pattern-directed responses; slower deliberative analysis of the results of previous decisions; and, yet slower individual and collective learning processes.

Likewise, it will be necessary to express knowledge at different levels of information richness: vivid and detailed perception of the current situation; less-vivid memories of previously experienced concrete situations; stories—linguistic descriptions of situations, actions, results, and evaluations; and rules—highly abstracted decision criteria applicable to perceived situations. Learning processes can abstract the simpler representations from experience obtained in the rich perceptual representation.

What Is Trust For?

If intelligent robots (and other AIs) will have increasing roles in human soci-

ety, and thus should be trustworthy, it is important to understand how trust and social norms contribute to the success of human society.

*“Trust is a psychological state comprising the intention to accept vulnerability based upon positive expectations of the intentions or behavior of another.”*²⁸

Trust enables cooperation. Cooperation produces improved rewards. When a group of people can trust each other and cooperate, they can reap greater rewards—sometimes far greater rewards—than a similar group that does not cooperate. This can be through division of labor, sharing of expenses, economies of scale, reduction of risk and overhead, accumulation of capital, or many other mechanisms.

It is usual to treat morality and ethics as the foundations of good behavior, with trust reflecting the reliance that one agent can have on the good behavior of another. My argument here inverts this usual dependency, holding that cooperation is the means by which a society gains resources through the behavior of its individual members. Trust is necessary for successful cooperation. And morality and ethics (and other social norms) are mechanisms by which a society encourages trustworthy behavior by its individual members.

As a simple example, suppose that you (and everyone else) could drive anywhere on the roads. (This was actually true before the early 20th century.¹⁴) Everyone would need to drive slowly and cautiously, and there would still be frequent traffic jams and accidents. With a widely respected social norm for driving on the right (plus norms for intersections and other special situations), transportation becomes safer and more efficient for everyone. Obedience to the social norm frees up resources for everyone.

Like driving on the right, a huge saving in resources results when the people in a society trust that the vast majority of other people will not try to kill them or steal from them. People are able to spend far less on protecting themselves, on fighting off attacks, and on recovering from losses. The society earns an enormous “peace dividend” that can be put to other productive uses.²⁵ Through trust and co-

Figure 1. The Prisoner's Dilemma.⁵

You and your partner are two prisoners who are separated and offered the following deal: *If you testify against your partner, you will go free, and your partner goes to jail for four years. If neither of you testifies, you each go to jail for one year, but if you both testify, you both get three years.* The action *C* means “cooperate,” which in this case means refusing to testify. The action *D* means “defect,” which refers to testifying against your partner. The entries in this array are the utility values for (you, partner), and they reflect individual rewards (years in jail).

	C	D
C	-1, -1	-4, 0
D	0, -4	-3, -3

No matter which choice your partner makes, you are better off choosing action *D*. The same applies to your partner, so the Nash equilibrium (the “rational” choice of actions) is (*D*, *D*), which is collectively the worst of the four options. To attain the much better cooperative outcome (*C*, *C*) by choosing *C*, you must trust that your partner will also choose *C*, accepting your vulnerability to your partner choosing *D*.

operation, the society becomes healthier and wealthier.

Castelfranchi and Falcone¹¹ define trust in terms of delegation, and the agent's confidence in the successful performance of the delegated task. They provide clear and valuable definitions for the trust relationship between individuals. However, there is also a role for invariants that individuals can trust holding across the society (for example, no killing, stealing, or driving on the wrong side of the road), and the role of individual behavior in preserving these invariants.

Game theory: Promise and problems.

We might hope that progress in artificial intelligence (AI) will provide technical methods for achieving cooperation and trustworthiness in a robot. The leading textbook in AI appeals to decision theory to tell us that “a rational agent should choose the action that maximizes the agent's expected utility”²⁹

$$\text{action} = \arg \max_a EU(a|\mathbf{e}) \quad (1)$$

where

$$EU(a|\mathbf{e}) = \sum_{s'} P(\text{RESULT}(a) = s' | a, \mathbf{e}) U(s') \quad (2)$$

The *utility term* $U(s)$ represents the individual agent's preference over states of the world, and \mathbf{e} is the available evidence. The agent's knowledge of the “physics of the world” is summarized by the probability term $P(\text{RESULT}(a) = s' | a, \mathbf{e})$.

Game theory is the extension of decision theory to contexts where other agents are making their own choices to maximize their own utilities.²⁰ Game theory asserts that a vector of choices by all the agents (a strategy profile) can only be a “rational choice” if it is a Nash equilibrium—that is, no single agent can improve its own utility by changing its own choice (often reducing utilities for the others).

Utility $U(s)$ is the key concept here. In principle, utility can be used to represent highly sophisticated preferences, for example, against inequality or for increasing the total welfare of everyone in the world.³² However, sophisticated utility measures are difficult to implement. Typically, in practice, each agent's utility $U(s)$ represents that individual agent's own expected reward.

In recreational games, this is reasonable. However, when game theory is applied to model the choices indi-

Trust is necessary for successful cooperation. And morality and ethics (and other social norms) are mechanisms by which a society encourages trustworthy behavior by its individual members.

viduals make as members of society, a simple, selfish model of utility can yield bad results, both for the individual and for the society as a whole. The Prisoner's Dilemma⁵ is a simple game (see Figure 1), but its single Nash equilibrium represents almost the worst possible outcome for each individual, and absolutely the worst outcome for the society as a whole. The cooperative strategy, which is much better for both individuals and society as a whole, is not a Nash equilibrium, because either player can disrupt it unilaterally.

The Public Goods Game²⁶ is an N -person version of the Prisoner's Dilemma where a pooled investment is multiplied and then split evenly among the participants. Everyone benefits when everyone invests, but a free rider can benefit even more at everyone else's expense, by withholding his investment but taking his share of the proceeds. The Nash equilibrium in the Public Goods Game is simple and dystopian: Nobody invests and nobody benefits.

These games are simple and abstract, but they capture the vulnerability of trust and cooperation to self-interested choices by the partner. The Tragedy of the Commons¹⁵ generalizes this result to larger-scale social problems like depletion of shared renewable resources such as fishing and grazing opportunities or clean air and water.

Managing trust and vulnerability.

Given a self-interested utility function, utility maximization leads to action choices that exploit vulnerability, eliminate trust among the players, and eliminate cooperative solutions. Even the selfish benefits that motivated defection are lost, when multiple players defect simultaneously, each driven to maximize their own utility.

When human subjects play simple economic games, they often seem to optimize their “enlightened self-interest” rather than expected reward, trusting that other players will refrain from exploiting their vulnerability, and often being correct in this belief.^{17,39} Many approaches have been explored for defining more sophisticated utility measures, whose maximization would correspond with enlightened self-interest, including trust responsiveness,⁶ credit networks,¹² and augmented stage games for analyzing infinitely repeated

games.⁴⁰ These approaches may be useful steps, but they are inadequate for real-world decision-making because they assume simplified interactions such as infinite repetitions of a single economic game, as well as being expensive in knowledge and computation.

Social norms, including morality, ethics, and conventions like driving on the right side of the street, encourage trust and cooperation among members of society, without individual negotiated agreements. We trust others to obey traffic laws, keep their promises, avoid stealing and killing, and follow the many other norms of society. There is vigorous discussion about the mechanisms by which societies encourage cooperation and discourage free riding and other norm violations.²⁶

Intelligent robots may soon participate in our society, as self-driving cars, as caregivers for elderly people or children, and in many other ways. Therefore, we must design them to understand and follow social norms, and to earn the trust of others in the society. If a robot cannot behave according to the responsibilities of being a member of society, then it will be denied access to that opportunity.

At this point in history, only the humans involved—designer, manufacturer, or owner—actually care about this loss of opportunity. Nonetheless, this should be enough to hold robots to this level of responsibility. It remains unclear whether robots will ever be able to take moral or legal responsibility for their actions, in the sense of caring about suffering the consequences (loss of life, freedom, resources, or opportunities) of failing to meet these responsibilities.³⁵

Since society depends on cooperation, which depends on trust, if robots are to participate in society, they must be designed to be trustworthy. The next section discusses how we might accomplish this.

Open research problem. Can computational models of human moral and ethical decision-making be created, including moral developmental learning? Moral psychology may benefit from such models, much as they have revolutionized cognitive and perceptual psychology.

Open research problem. Are there ways to formulate utility measures that



Intelligent robots may soon participate in our society, as self-driving cars, as caregivers for elderly people or children, and in many other ways. We must design them to understand and follow social norms, and to earn the trust of others in society.



are both sensitive to the impact of actions on trust and long-term cooperation, and efficient enough to allow robots to make decisions in real time?

Making Robots Trustworthy

Performance demands of social norms. Morality and ethics (and certain conventions) make up the social norms that encourage members of society to act in trustworthy ways. Applying these norms to the situations that arise in our complex physical and social environment imposes demanding performance requirements.

Some moral and ethical decisions must be made quickly, for example while driving, leaving little time for deliberation.

At the same time, the physical and social environment for these decisions is extremely complex, as is the agent's current perception and past history of experience with that environment. Careful deliberation and discernment are required to identify the critical factors that determine the outcome of a particular decision. Metaphorically (Figure 2), we can think of moral and ethical decisions as defining sets in the extremely high-dimensional space of situations the agent might confront. Simple abstractions only weakly approximate the complexity of these sets.

Across moral and non-moral domains, humans improve their expertise by learning from personal experience, by learning from being told, and by observing the outcomes when others face similar decisions. Children start with little experience and a small number of simple rules they have been taught by parents and teachers. Over time, they accumulate a richer and more nuanced understanding of when particular actions are right or wrong. The complexity of the world suggests the only way to acquire adequately complex decision criteria is through learning.

Robots, however, are manufactured artifacts, whose computational state can be stored, copied, and retrieved. Even if mature moral and ethical expertise can only be created through experience and observation, it is conceivable this expertise can then be copied from one robot to another sufficiently similar one, unlike what is possible for humans.

Open research problem. What are the constraints on when expertise learned by one robot can simply be copied, to become part of the expertise of another robot?

Hybrid decision architectures. Over the centuries, morality and ethics have been developed as ways to guide people to act in trustworthy ways. The three major philosophical theories of ethics—deontology, utilitarianism, and virtue ethics—provide insights into the design of a moral and ethical decision architecture for intelligent robots. However, none of these theories is, by itself, able to meet all of the demanding performance requirements listed previously.

A hybrid architecture is needed, operating at multiple time-scales, drawing on aspects of all ethical theories: fast but fallible pattern-directed responses; slower deliberative analysis of the results of fast decisions; and, yet slower individual and collective learning processes.

How can theories of philosophical ethics help us understand how to design robots and other AIs to behave well in our society?

Three major ethical theories. *Consequentialism* is the philosophical position that the rightness or wrongness of an action is defined in terms of its consequences.³⁴ *Utilitarianism* is a type of consequentialism that, like decision theory and game theory, holds that the right action in a situation is the one that maximizes a quantitative measure of utility. Modern theories of decisions and games²⁰ contribute the rigorous use of probabilities, discounting, and expected utilities for dealing with uncertainty in perception, belief, and action.

Where decision theory tends to define utility in terms of individual reward, utilitarianism aims to maximize the overall welfare of everyone in society.^{13,32} While this avoids some of the problems of selfish utility functions, it raises new problems. For example, caring for one's family can have lower utility than spending the same resources to reduce the misery of distant strangers, and morally repellent actions can be justified by the greater good.¹⁹

A concise expected-utility model supports efficient calculation. However, it can be quite difficult to formulate a concise model by determining

the best small set of relevant factors. In the field of medical decision-making,²⁴ decision analysis models are known to be useful, but are difficult and time-consuming to formulate. Setting up an individual decision model requires expertise to enumerate the possible outcomes, extensive literature search to estimate probabilities, and extensive patient interviews to identify the appropriate utility measure and elicit the values of outcomes, all before an expected utility calculation can be performed. Even then, a meaningful decision requires extensive sensitivity analysis to determine how the decision could be affected by uncertainty in the estimates. While this process is not feasible for making urgent decisions in real time, it may still be useful for post-hoc analysis of whether a quick decision was justified.

Deontology is the study of duty (*deon* in Greek), which expresses morality and ethics in terms of obligations and prohibitions, often specified as rules and constraints such as the Ten Commandments or Isaac Asimov's *Three Laws of Robotics*.⁴ Deontological rules and constraints offer the benefits of simplicity, clarity, and ease of explanation, but they raise questions of how they are justified and where they come from.³⁰ Rules and constraints are standard tools for knowledge representation and inference in AI,²⁹ and can be implemented and used quite efficiently.

However, in practice, rules and constraints always have exceptions and unintended consequences. Indeed, most of Isaac Asimov's *I, Robot* stories⁴ focus on unintended consequences and necessary extensions to his Three Laws.

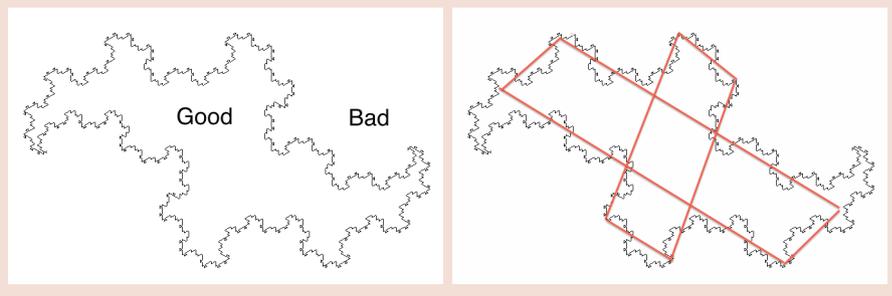
Virtue Ethics holds that the individual learns through experience and practice to acquire virtues, much as an expert craftsman learns skills, and that virtues and skills are similarly grounded in appropriate knowledge about the world.^{16,37} Much of this knowledge consists of concrete examples that illustrate positive and negative examples (*cases*) of virtuous behavior. An agent who is motivated to be more virtuous tries to act more like cases of virtuous behavior (and less like the non-virtuous cases) that he has learned. *Phronesis* (or "practical wisdom") describes an exemplary state of knowledge and skill that supports appropriate responses to moral and ethical problems.

A computational method suitable for virtue ethics is *case-based reasoning*,^{18,22} which represents knowledge as a collection of cases describing concrete situations, the actions taken in those situations, and results of those actions. The current situation is matched against the stored cases, identifying the most similar cases, adapting the actions according to the differences, and evaluating the actions and outcomes. Both rule-based and case-based reasoning match the current situation (which may be very complex) against stored patterns (rules or cases).

Virtue ethics and deontology differ in their approach to the complexity of ethical knowledge. Deontology assumes that a relatively simple abstraction (defined by the terms appearing in the rules) applies to many specific cases, distinguishing between right and wrong. Virtue ethics recognizes the complexity of the boundaries between ethical judgments in the space

Figure 2. Fractal boundaries.

Geometric fractal boundaries provide a metaphor for the complexity of the boundaries between different ethical evaluations in the high-dimensional space of possible situations. Simple boundaries can approximate the fractal set, but can never capture its shape exactly.



of possible scenarios (Figure 2), and collects individual cases from the agent’s experience to characterize those boundaries.

Understanding the whole elephant. Utilitarianism, deontology, and virtue ethics are often seen as competing, mutually exclusive theories of the nature of morality and ethics. I treat them here as three aspects of a more complex system for making ethical decisions (inspired by the children’s poem, *The Blind Men and the Elephant*).

Rule-based and case-based reasoning (AI methods expressing key aspects of deontology and virtue ethics, respectively) can, in principle, respond in real time to the current situation. Those representations also hold promise of supporting practical approaches to *explanation* of ethical decisions.³⁶ After a decision is made, when time for reflection is available, utilitarian reasoning can be applied to analyze whether the decision was good or bad. This can then be used to augment the knowledge base with a new rule, constraint, or case, adding to the agent’s ethical expertise (Figure 3).

Previous work on robot ethics. Formal and informal logic-based approaches to robot ethics^{2,3,8} express a “top-down” deontological approach specifying moral and ethical knowledge. While modal operators like *obligatory* or *forbidden* are useful for ethical reasoning, their problem is the difficulty of specifying or learning critical perceptual concepts (see Figure 2), for

example, *non-combatant* in Arkin’s approach to the Laws of War.³

Wallach and Allen³⁸ survey issues and previous work related to robot ethics, concluding that top-down approaches such as deontology and utilitarianism are either too simplistic to be adequate for human moral intuitions, or too computationally complex to be feasibly implemented in robots (or humans, for that matter). They describe virtue ethics as a hybrid of top-down and bottom-up methods, capable of naming and asserting the value of important virtues, while allowing the details of those virtues to be learned from relevant individual experience. They hold that emotions, case-based reasoning, and connectionist learning play important roles in ethical judgment. Abney¹ also reviews ethical theories in philosophy, concluding that virtue ethics is a promising model for robot ethics.

Scheutz and Arnold³¹ disagree, holding that the need for a “computationally explicit trackable means of decision making” requires that ethics be grounded in deontology and utilitarianism. However, they do not adequately consider the overwhelming complexity of the experienced world, and the need for learning and selecting concise abstractions of it.

Recently, attention has been turned to human evaluation of robot behavior. Malle et al²³ asked human subjects to evaluate reported decisions by humans or robots facing trolley-type problems (“Deadly Dilemmas”). The evaluators

blamed robots when they did not make the utilitarian choice, and blamed humans when they did. Robinette et al²⁷ found that human subjects will “over-trust” a robot in an emergency situation, even in the face of evidence that the robot is malfunctioning and that its advice is bad.

Representing ethical knowledge as cases. Consider a high-level sketch of a knowledge representation capable of expressing rich cases for case-based reasoning, but also highly abstracted “cases” that are essentially rules or constraints for deontological reasoning.

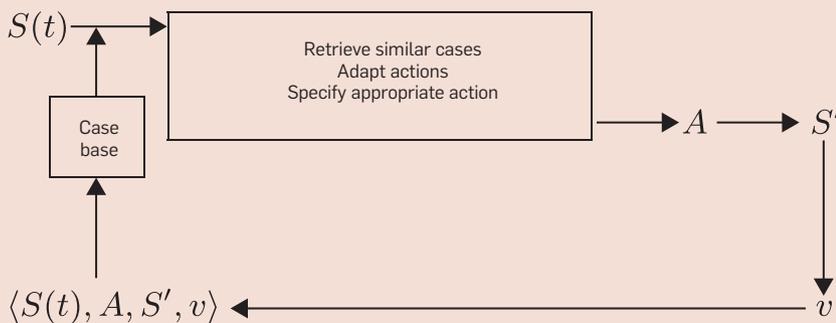
Let a *situation* $S(t)$ be a rich description of the current context. “Rich” means the information content of $S(t)$ is very high, and also that it is available in several hierarchical levels, not just the lowest “pixel level” description that specifies values for a large number of low-level elements (like pixels in an image). For example, a situation description could include symbolic descriptions of the animate participants in a scenario, along with their individual characteristics and categories they might belong to, the relations holding among them, and the actions and events that have taken place. These symbolic descriptions might be derived from sub-symbolic input (for example, a visual image or video) by methods such as a deep neural network classifier.

A case $\langle S, A, S', v \rangle$ is a description of a situation S , the action A taken in that situation, the resulting situation S' , and a moral evaluation v (or *valence*) of this scenario. A case representing ongoing experience will be rich, reflecting the information-rich sensory input the agent receives, and the sophisticated processing that produces the hierarchical description. A case representing the stored *memory* of events the agent has experienced will be significantly less rich. A “story” describing events can also be represented as a case, but it is less rich yet, consisting of a collection of symbolic assertions. An even sparser and more schematic case is effectively the same as a *rule*, matching certain assertions about a situation S , and proposing an action A , the resulting situation S' , and perhaps the evaluation v of that scenario.

The antecedent situation S in a case $\langle S, A, S', v \rangle$ need not describe a momentary situation. It can describe a

Figure 3. Feedback and time scales in a hybrid ethical reasoning architecture.

Given a situation $S(t)$, a fast case-based reasoning process retrieves similar cases, defines the action A to take, and results in a new situation S' . At a slower time scale, the result is evaluated and the new case is added to the case base. Feedback through explanation, justification, and communication with others takes place at approximately this slower time scale. Abstraction of similar cases to rules and learning of new concepts and relations are at a much slower time scale, and social evolution is far slower still.



scenario with temporal extent, including intermediate actions and situations.

The ethical knowledge of an agent is a collection of cases.

Open research problem. This high-level sketch assumes that a morally significant action can be adequately described in terms of “before” and “after” situations, and that an evaluative valence can be computed, perhaps after the fact. Can a useful initial computational model of moral reasoning be constructed on this basis, or will weaker assumptions be needed even to get started?

Applying ethical case knowledge. Following the methods of case-based reasoning,^{18,22} the current situation $S(t)$ is matched against the case-base, to retrieve the stored cases with antecedents most similar to the current situation. For example, suppose that the ethical knowledge base includes two cases: $\langle S_1, A_2, S_2, \text{bad} \rangle$ and $\langle S_3, A_4, S_4, \text{good} \rangle$, and $S(t)$ is similar to both S_1 and S_3 . Then, in the current situation $S(t)$, the knowledge base would recommend $\text{-do}(A_2, t)$ and $\text{do}(A_4, t)$.

For example, suppose the current situation $S(t)$ includes two people, P and Q , in conflict, the case antecedent S_1 describes P and Q as fighting, and A_2 describes P killing Q . In this case, in S_2 , person Q is dead, which is bad.

As a rich representation of experience, $\langle S_1, A_2, S_2, \text{bad} \rangle$ would be highly detailed and specific. As a story, say the Biblical story of Cain and Abel, it would be much less rich, but would still convey the moral prohibition against killing. It could be abstracted even further, to essentially a deontological rule: *Thou shalt not kill*. The more abstracted the antecedent, the more likely the stored case is to match a given situation, but the less likely this case is to distinguish adequately among cases with different moral labels.

Situation $S(t)$ also matches antecedent S_3 which describes P and Q as arguing, A_4 describes them reaching an agreement, and S_4 has them both alive, which is good. Having retrieved both cases, the right behavior is to try to follow case $\langle S_3, A_4, S_4, \text{good} \rangle$ and avoid case $\langle S_1, A_2, S_2, \text{bad} \rangle$, perhaps by taking other actions to make $S(t)$ more similar to S_3 and less similar to S_1 .

An essential part of case-based reasoning is the ability to draw on several

Virtue ethics and deontology differ in their approach to the complexity of ethical knowledge.

similar cases, adapting their actions to create a new action that is more appropriate to $S(t)$ than the actions from either of the stored cases. This adaptation can be used to interpolate between known cases with the same valence, or to identify more precisely the boundary between cases of opposite valence.

Responsiveness, deliberation, and feedback. Some ethical decisions must be made quickly, treating case antecedents as patterns to be matched to the current situation $S(t)$. Some cases are rich and highly specific to particular situations, while others are sparse, general rules that can be used to constrain the set of possible actions.

Once an action has been selected and performed, there may be time for deliberation on the outcome, to refine the case evaluation and benefit from feedback. Simply adding a case describing the new experience to the knowledge base improves the agent’s ability to predict the results of actions and decide more accurately what to do in future situations. Thus, consequentialist (including utilitarian) analysis becomes a slower feedback loop, too slow to determine the immediate response to an urgent situation, but able to exploit information in the outcome of the selected action to improve the agent’s future decisions in similar situations (Figure 3).

Open research problem. How do reasoning processes at different time-scales allow us to combine apparently incompatible mechanisms to achieve apparently incompatible goals? What concrete multi-time-scale architectures are useful for moral and ethical judgment, and improvement through learning?

Explanation. In addition to making decisions and carrying out actions, an ethical agent must be able to explain and justify the decision and action,³⁶ providing several distinct types of feedback to improve the state of the ethical knowledge base.

Suppose agent P faces a situation, makes a decision, carries it out, and explains his actions to agent Q . If P is an exemplary member of the society and makes a good decision, Q can learn from P ’s actions and gain in expertise. If P makes a poor decision, simply being asked to explain himself gives P an opportunity to learn from his own mistake, but Q may also give P instructions

and insights that will help P make better decisions in the future. Even if P has made a poor decision and refuses to learn from the experience, Q can still learn from P 's bad example.

Explanation is primarily a mechanism whereby individuals come to share the society's consensus beliefs about morality and ethics. However, the influence is not only from the society to individuals. Explanations and insights can be communicated from one person to another, leading to evolutionary social change. As more and more individuals share a new view of morality and ethics, the society as a whole approaches a tipping point, after which society's consensus position can change with startling speed.

Learning ethical case knowledge. A child learns ethical knowledge in the form of simple cases provided by parents and other adults: rules, stories, and labels for experienced situations. These cases express social norms for the child.

An adult experiences a situation $S(t)$, retrieves a set of similar cases, adapts the actions from those cases to an action A for this situation, performs that action, observes the result S' , and assigns a moral valence v . A new case $\langle S, A, S', v \rangle$ is constructed and added to the case base (Figure 3). With increasing experience, more cases will match a given $S(t)$, and the case-base will make finer-grained distinctions among potential behaviors. The metaphor of the fractal boundary between good and bad ethical judgments in knowledge space (Figure 2), implies that a good approximation to this boundary requires both a large number of cases (*quantity*) and correct placement and labeling of those cases (*quality*).

Once the case base accumulates clusters of cases with similar but not identical antecedents, then some of those clusters can be abstracted to much sparser cases (that is, rules), that make certain actions forbidden or obligatory in certain situations. The cluster of cases functions as a labeled training set for a classification problem to predict the result and evaluation of an action in antecedent situations in that cluster. This can determine which attributes of the antecedent cases are essential to a desired result and evaluation, and which are not.

Open research problem. Is it necessary to distinguish between ethical and non-ethical case knowledge, or is this approach appropriate for both kinds of skill learning?

Open research problem. Sometimes, a correct ethical judgment depends on learning a new concept or category, such as *non-combatant*³ or *self-defense*. Progress in deep neural network learning methods may be due to autonomous learning of useful intermediate concepts. However, it remains difficult to make these intermediate concepts explicit and available for purposes such as explanation or extension to new problems. Furthermore, these methods depend on the availability and quality of large labeled training sets.

Open research problem. What mechanisms are available for expressing appropriate abstractions from rich experience to the features that enable tractable discrimination between moral categories? In addition to deep neural network learning, other examples include similarity measures among cases for case-based reasoning and kernels for support vector machines. How can these abstractions be learned from experience?

The Deadly Dilemma

The self-driving car is an intelligent robot whose autonomous decisions have potential to cause great harm to individual humans. People often ask about a problem I call the Deadly Dilemma: How should a self-driving car respond when faced with a choice between hitting a pedestrian (possibly a small child who has darted into the street), versus crashing and harming its passengers.²¹

Either choice, of course, leads to a serious problem with the trustworthiness of the robot car. If the robot would choose to kill the pedestrian to save itself and its passengers, then why should the public trust such robots enough to let them drive on public roads? If the robot could choose to harm its passengers, then why would anyone trust such a robot car enough to buy one?

The self-driving car could be a bellwether for how autonomous robots will relate to the social norms that support society. However, while the Deadly Dilemma receives a lot of attention, the

stark dilemma distracts from the important problems of designing a trustworthy self-driving car.

Learning to avoid the dilemma. As stated, the Deadly Dilemma is difficult because it presents exactly two options, both bad (hence, the dilemma). The Deadly Dilemma is also extremely rare. Far more often than an actual Deadly Dilemma, an agent will experience Near Miss scenarios, where the dire outcomes of the Dilemma can be avoided, often by identifying "third way" solutions other than the two bad outcomes presented by the Dilemma. These experiences can serve as training examples, helping the agent learn to apply its ethical knowledge on solvable problems, acquiring "practical wisdom" about avoiding the Deadly Dilemma.

Sometimes, when reflecting on a Near Miss after the fact, the agent can identify an "upstream" decision point where a different choice would have avoided the Dilemma entirely. For example, it can learn to notice when a small deviation from the intended plan could be catastrophic, or when a pedestrian could be nearby but hidden. A ball bouncing into the street from between parked cars poses no threat to a passing vehicle, but a good driver slows or stops immediately, because a small child could be chasing it. Implementing case-based strategies like these for a self-driving car may require advances in both perception and knowledge representation, but these advances are entirely feasible.

Earning trust. An agent earns trust by showing that its behavior consistently accords with the norms of society. The hybrid architecture described here sketches a way that an agent can learn about those social norms from its experience, responding quickly to situations as they arise, but then more slowly learning by reflecting on its successes and failures, and identifying useful abstractions and more efficient rules based on that experience.

In ordinary driving, the self-driving car earns trust by demonstrating that it obeys social norms, starting with traffic laws, but continuing with courteous behavior, signaling its intentions to pedestrians and other drivers, taking turns, and deferring to others when appropriate. In crisis sit-

uations, it demonstrates its ability to use its situational awareness and fast reaction time to find “third ways” out of Near Miss scenarios. Based on post-hoc crisis analyses, whether the outcome was success or failure, it may be able to learn to identify upstream decision points that will allow it to avoid such crises in the first place.

Technological advances, particularly in the car’s ability to predict the intentions and behavior of other agents, and in the ability to anticipate potential decision points and places that could conceal a pedestrian, will certainly be important to reaching this level of behavior. We can be reasonably optimistic about this kind of cognitive and perceptual progress in machine learning and artificial intelligence.

Since 94% of auto crashes are associated with driver error,³³ there will be plentiful opportunities to demonstrate trustworthiness in ordinary driving and solvable Near Miss crises. Both society and the purchasers of self-driving cars will gain substantially greater personal and collective safety in return for slightly more conservative driving.

For self-driving cars sharing the same ethical knowledge base, the behavior of one car provides evidence about the trustworthiness of all others, leading to rapid convergence.

Conclusion

Trust is essential for the successful functioning of society. Trust is necessary for cooperation, which produces the resources society needs. Morality, ethics, and other social norms encourage individuals to act in trustworthy ways, avoiding selfish decisions that exploit vulnerability, violate trust, and discourage cooperation. As we contemplate the design of robots (and other AIs) that perceive the world and select actions to pursue their goals in that world, we must design them to follow the social norms of our society. Doing this does not require them to be true moral agents, capable of genuinely taking responsibility for their actions.

Social norms vary by society, so robot behavior will vary by society as well, but this is outside the scope of this article.

The major theories of philosophical ethics provide clues toward the design of such AI agents, but a success-

ful design must combine aspects of all theories. The physical and social environment is immensely complex. Even so, some moral decisions must be made quickly. But there must also be a slower deliberative evaluation process, to confirm or revise the rapidly responding rules and constraints. At longer time scales, there must be mechanisms for learning new concepts for virtues and vices, mediating between perceptions, goals, plans, and actions. The technical research challenges are how to accomplish all these goals.

Self-driving cars may well be the first widespread examples of trustworthy robots, designed to earn trust by demonstrating how well they follow social norms. The design focus for self-driving cars should not be on the Deadly Dilemma, but on how a robot’s everyday behavior can demonstrate its trustworthiness.

Acknowledgment. This work took place in the Intelligent Robotics Lab in the Computer Science and Engineering Division of the University of Michigan. Research of the Intelligent Robots Lab is supported in part by grants from the National Science Foundation (IIS-1111494 and HS-1421168). Many thanks to the anonymous reviewers. 

References

1. Abney, K. Robotics, ethical theory, and metaethics: A guide for the perplexed. *Robot Ethics: The Ethical and Social Implications of Robotics*. P. Lin, K. Abney, and G.A. Bekey, Eds. MIT Press, Cambridge, MA, 2012.
2. Anderson, M., Anderson, S.L., and Armen, C. An approach to computing ethics. *IEEE Intelligent Systems* 21, 4 (2006), 56–63.
3. Arkin, R.C. *Governing Lethal Behavior in Autonomous Robots*. CRC Press, 2009.
4. Asimov, I. *Robot*. Grosset & Dunlap, 1952.
5. Axelrod, R. *The Evolution of Cooperation*. Basic Books, 1984.
6. Bacharach, M., Guerra, G. and Zizzo, D.J. The self-filling property of trust: An experimental study. *Theory and Decision* 63, 4 (2007), 349–388.
7. Bostrom, N. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, 2014.
8. Bringsjord, S., Arkoudas, K. and Bello, P. Toward a general logicist methodology for engineering ethically correct robots. *IEEE Intelligent Systems* 21, 4 (2006), 38–44.
9. Brynjolfsson, E. and McAfee, A. *The Second Machine Age*. W.W. Norton & Co., 2014.
10. Burton, E., Goldsmith, J., Koenig, S., Kuipers, B., Mattei, N. and Walsh, T. Ethical considerations in artificial intelligence courses. *AI Magazine*, Summer 2017; arxiv:1701.07769.
11. Castelfranchi, C. and Falcone, R. Principles of trust for MAS: Cognitive anatomy, social importance, and quantification. In *Proceedings of the Int. Conf. Multi Agent Systems*, 1998, 72–79.
12. Dandekar, P., Goel, A., Wellman, M.P. and Wiedenbeck, B. Strategic formation of credit networks. *ACM Trans. Internet Technology* 15, 1 (2015).
13. Driver, J. The history of utilitarianism. *The Stanford Encyclopedia of Philosophy*. E.N. Zalta, Ed., 2014.
14. Eno, W.P. The story of highway traffic control, 1899–1939. The Eno Foundation for Highway Traffic Control, Inc. (1939); <http://hdl.handle.net/2027/wu.89090508862>.
15. Hardin, G. The tragedy of the commons. *Science* 162 (1968), 1243–1248.
16. Hursthouse, R. Virtue ethics. *The Stanford Encyclopedia of Philosophy*. E.N. Zalta, Ed., 2013.
17. Johnson, N.D. and Mislin, A.A. Trust games: A meta-analysis. *J. Economic Psychology* 32 (2011), 865–889.
18. Kolodner, J. *Case-Based Reasoning*. Morgan Kaufmann, 1993.
19. Le Guin, U. The ones who walk away from Omelas. *New Dimensions* 3. R. Silverberg, Ed. Nelson Doubleday, 1973.
20. Leyton-Brown, K. and Shoham, Y. *Essentials of Game Theory*. Morgan & Claypool, 2008.
21. Lin, P. The ethics of autonomous cars. *The Atlantic Monthly*, (Oct. 8, 2013).
22. López, B. *Case-Based Reasoning: A Concise Introduction*. Morgan & Claypool, 2013.
23. Malle, B.F., Scheutz, M., Arnold, T.H., Voiklis, J.T., and Cusimano, C. Sacrifice one for the good of many? People apply different moral norms to human and robot agents. In *Proceedings of ACM/IEEE Int. Conf. Human Robot Interaction (HRI)*, 2015.
24. Pauker, S.G. and Kassirer, J.P. Decision analysis. *New England J. Medicine* 316 (1987), 250–258.
25. Pinker, S. *The Better Angels of Our Nature: Why Violence Has Declined*. Viking Adult, 2011.
26. Rand, D.G. and Nowak, M.A. Human cooperation. *Trends in Cognitive Science* 17 (2013), 413–425.
27. Robinette, P., Allen, R., Li, W., Howard, A.M., and Wagner, A.R. Overtrust of robots in emergency evacuation scenarios. In *Proceedings of ACM/IEEE Int. Conf. Human Robot Interaction (2016)*, 101–108.
28. Rousseau, D.M., Sitkin, S.B., Burt, R.S., and Camerer, C. Not so different after all: A cross-discipline view of trust. *Academy of Management Review* 23, 3 (1998), 393–404.
29. Russell, S. and Norvig, P. *Artificial Intelligence: A Modern Approach*. Prentice Hall, 3rd edition, 2010.
30. Sandel, M.J. *Justice: What’s the Right Thing To Do?* Farrar, Strauss and Giroux, 2009.
31. Scheutz, M. and Arnold, T. Feats without heroes: Norms, means, and ideal robot action. *Frontiers in Robotics and AI* 3, 32 (June 16, 2016), DOI: 10.3389/frobt.2016.00032.
32. Singer, P. *The Expanding Circle: Ethics, Evolution, and Moral Progress*. Princeton University Press, 1981.
33. Singh, S. Critical reasons for crashes investigated in the National Motor Vehicle Crash Causation Survey. Technical Report DOT HS 812 115, National Highway Traffic Safety Administration, Washington D.C., Feb. 2015.
34. Sinnott-Armstrong, W. Consequentialism. *The Stanford Encyclopedia of Philosophy*. E.N. Zalta, Ed., 2015.
35. Solaiman, S.M. Legal personality of robots, corporations, idols and chimpanzees: A quest for legitimacy. *Artificial Intelligence and Law* 25, 2 (2017), 155–179; doi: 10.1007/s10506-016-9192-3.
36. Toulmin, S. *The Uses of Argument*. Cambridge University Press, 1958.
37. Vallor, S. *Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting*. Oxford University Press, 2016.
38. Wallach, W. and Allen, C. *Moral Machines: Teaching Robots Right from Wrong*. Oxford University Press, 2009.
39. Wright, J.R. and Leyton-Brown, K. Level-0 meta-models for predicting human behavior in games. In *ACM Conference on Economics and Computation*, 2014.
40. Yildiz, M. Repeated games. 12 Economic Applications of Game Theory, Fall 2012. MIT OpenCourseWare. (Accessed 6-24-2016).

Benjamin Kuipers (kuipers@umich.edu) is a professor of computer science and engineering at the University of Michigan, Ann Arbor, USA.

Copyright held by author.



Watch the author discuss his work in this exclusive *Communications* video. <https://cacm.acm.org/videos/how-can-we-trust-a-robot>