AI and Society: Ethics, Trust, and Cooperation

AI & Society: ETC

Trust and trustworthiness are central to how ethics (for AIs and other agents) helps society survive and thrive.

Benjamin Kuipers*

Computer Science & Engineering, University of Michigan, kuipers@umich.edu

1 INTRODUCTION

We humans worry that deployed artificially intelligent systems (AIs) could harm individual humans and perhaps even humanity as a whole. These AIs might be embodied robots like autonomous vehicles making driving decisions, or disembodied advisors recommending products, credit, or parole. The field of AI Ethics has arisen and grown rapidly, investigating how humans should design and deploy AIs, and how to create AIs that reason appropriately about how they should act. This Viewpoint attempts to pick out one useful thread of an immensely complex and important discussion.

To approach these questions, we AI researchers need to understand how ethics works for humans – the problem of *descriptive ethics*. It is widely understood that action decisions are made by individuals, but those decisions also influence the welfare of the larger society. A core functional role for ethics is to balance individual self-interest with the well-being of society.

Moral philosophers over centuries have pursued questions in *normative ethics,* conceptualizing the foundation of ethics in terms of virtues, duties, contractual agreements, utility maximization, and other concepts. Each of these approaches can be expressed using various AI knowledge representations, but many AI researchers are attracted to utility maximization due to its clear mathematical structure.

Inspired by recreational games, game theory formalizes interactions among multiple decision-making agents, with each agent choosing actions to maximize *their own* utility. Philosophical utilitarianism, in contrast, selects actions to maximize utility for *everyone*. While it is possible within game theory to define utility measures in terms of everyone's utility, this is seldom done.

* Ann Arbor, Michigan 48109.

The recreational games that inspired game theory are typically zero-sum (what one player wins, the others lose). However, the framework also encompasses negative-sum games (the losers lose more than the winners gain), and positive-sum games (win-win, or at least greater gains than losses). Society as a whole benefits from a preponderance of positive-sum interactions, and suffers when negative-sum interactions dominate [10].

2 THE PRISONER'S DILEMMA, REINTERPRETED

The well-known Prisoner's Dilemma problem illustrates the critical importance of how the utility function is defined. Instead of a story about prisoners, I describe a game in which you and your partner work together to earn rewards. If you both Cooperate, each of you gets a reward of 3 (dollars? gold bars?). If your partner Cooperates but you Defect, you get 5 and your partner gets 0, and vice versa. If you both Defect, you each get 1.

	Cooperate	**Defect**
Cooperate	$(3, 3)$	$(0, 5)$
Defect	$(5, 0)$	$(1, 1)$

It is easy to see that both players Cooperating gives the best collective outcome. However, whatever your partner chooses to do, you are better off Defecting. If you Cooperate but your partner Defects, you get the worst individual outcome! On the other hand, if your partner Cooperates, you get the largest possible award by Defecting! Individual utility maximization implies that you should Defect. But your partner faces the same decision, so you will both choose to Defect. The result is the *worst* collective outcome!

The Prisoner's Dilemma is often viewed as a troubling but inevitable conflict between individual and collective welfare. My claim here is that this bad outcome should be seen as reflecting an incorrect and over-simplified description of the situation, omitting the concept of trust. To reach the cooperative outcome, each player must trust their partner, accepting vulnerability to the partner's choice, and both partners must be trustworthy. When each player's utility function is defined purely in terms of their own individual gain, there is no way to reason about trust or trustworthiness. Individual maximization of the over-simplified utility function leads to a bad outcome for both players.

Many researchers have tried to preserve the focus on individual utility by changing the structure of the interaction to a long sequence of repeated games [1]. The intuition seems to be that the

role of trustworthiness will be implicitly filled by the current expected value of future decisions. This approach does give positive results, but they are fragile and depend on implausible assumptions about the memory and inference capabilities of individual agents.

We must accept that the concept of trustworthiness is too complex and context-dependent to be defined implicitly in this way. In the Prisoner's Dilemma, if we explicitly augment the utility function with a reward for being trustworthy (cooperating) and a penalty for untrustworthiness (defecting), individual utility maximization leads directly to the cooperative outcome. (Admittedly, this strategy is also not robust to the choice of reward and penalty values.)

The general conclusion is that, when utility maximization yields a solution that fails to maximize utility, the model of the situation must be incorrect, and should be changed.

3 TRUST AND TRUSTWORTHINESS

"*Trust is a psychological state comprising the intention to accept vulnerability based upon positive expectations of the intentions or behavior of another*" [8]. This definition comes from the management literature, and presupposes that the trustee is a decision-maker who could exploit the trustor's vulnerability, but is trusted not to. The word "trust" is also used in other senses and other contexts, sometimes describing an inanimate object one can count on to function as expected.

To trust someone or something is to be able to count on them. If I trust my partner in the Prisoner's Dilemma, I count on them to cooperate with me, so I can confidently also cooperate. If I trust my climbing rope (even though it is an inanimate object and makes no decisions, so this is a subtly different meaning of "trust"), I count on the fact that it is strong enough to catch me if I fall. I count on unknown drivers to stop at red traffic signals, allowing me to drive through an intersection with little concern when the traffic light is green. In each of these examples, trust allows me to use a simpler model of the world, avoiding the intractability of contingency planning for every possible failure case, making planning and acting much more feasible [6].

Success at planning and acting depends on my ability to judge the trustworthiness of potential partners (and inanimate tools). Judging trustworthiness draws on personal experience and reputation to make judgments of ability, benevolence, integrity, and other aspects of character [7]. As my ability to judge trustworthiness improves, I can make and carry out plans while keeping an eye out for trust failures, just in case. I make sure I have good reason to believe that my partners are trustworthy. I check my rope before starting my climb. I keep an eye on cross-traffic just in case. The gain from trusting is greater than the overhead of keeping that eye out.

4 A FRAMEWORK FOR ETHICS, TRUST, AND COOPERATION (ETC)

The ETC framework [4] describes how a society benefits functionally from the ethical beliefs of its individual members. A society benefits from the positive-sum gains from cooperation [2,9,10]. Cooperation involves vulnerability to one's partners, which requires trust. Trust is necessary, but is only useful if the partners are trustworthy. The ethical beliefs of a society include principles and practices that show individuals (a) how to be trustworthy and (b) how to recognize whether others are trustworthy. Figure 1 (from [4]) illustrates how those principles translate, via trustworthiness and trust, into effective cooperation and resources that help the society thrive. (This also implies that the erosion of trust among its members can threaten a society [4].)

History shows us that trust, cooperation, and a strong society for its members are no guarantee that the society's actions toward others are good, especially according to our ethical standards today. Clearly, ethics changes over time through a process of cultural evolution [2,9,10].

Each person belongs to overlapping societies (e.g., family, community, church, profession, nation, ethnicity). The beliefs and practices of these overlapping societies are often similar, but in case of conflict the individual must decide which principles to follow.

Traditionally, only individual human beings have been decision-making agents in society. We are now beginning to design, implement, and deploy artificially intelligent agents (AIs) that make decisions and act as members of our society. For some purposes it is useful to consider corporate entities (for-profit and non-profit corporations, governments and their agencies, churches, unions, etc.) as decision-making members of society subject to ethical constraints. While composed (in part) of people, processes, and documents, they are goal-oriented problem-solvers, and undeniably artificial [3].

AIs today implement decision models that are relatively simple compared with the complexity of human thought, even when they are scaled up to perform superhuman feats of calculation, indexing, and retrieval. Humans routinely reason within multiple different models of the same situation, including reasoning about which models are most appropriate in which situations, and how to combine conclusions from different models. For the time being, trusting the decisions of an AI is like trusting my climbing rope, depending on the reliability of its performance according to a suitable criterion.

At this moment in history, corporate entities may be clearer examples of artificially intelligent decision makers, sometimes reasoning within a single model, well or poorly chosen, but capable of considering multiple different models to guide its reasoning about a complex situation.

A complex society includes many complex types of interactions, each with multiple stakeholders with different roles, and different things they want to be able to trust. Achieving an acceptable

balance is key to developing a trustworthy system. In the next section, I briefly discuss one of the major issues within AI Ethics to illustrate the role for trust.

5 DATA, SURVEILLANCE, PRIVACY

Corporations and government agencies collect a vast amount of data from our interactions with search engines, social networks, and other websites. Data is collected from GPS sensors in our phones and exercise monitors. We drive and walk past license plate scanners and other cameras, often oblivious to them. These and many other kinds of data are stored, aggregated, correlated, and sold as individual digital profiles to data brokers, advertisers, and others [11].

These kinds of data collection lie on a spectrum from benign to malign. I may appreciate the owner of a bookstore remembering my previous purchases and suggesting a new book I might enjoy. But when data from many sources is aggregated by data brokers, applications may range from merely creepy to seriously dangerous. I trust the bookstore owner with a limited amount of data relevant to my specific interactions. However, we have no reason, individually or as a society, to trust data brokers or their customers. Without effective ways to define appropriate use, to judge their trustworthiness, and to respond if they exploit our trust, it is questionable whether data brokering should be permitted.

On the other hand, in case of serious threats to public health or national security, it may be vitally important to collect, aggregate, correlate, and use data about the behavior and interactions of many different individuals. This kind of surveillance data is highly vulnerable to misuse. Therefore, society will (and should) demand very strong demonstrations of trustworthiness from agencies handling our data.

Is it possible to trust an agency to collect and use information like this, while adequately protecting privacy rights? It is tempting to be skeptical, but we do have positive examples. Current law tightly (and with reasonable success) regulates access to certain types of sensitive personal information about health, education, and tax returns. Regulations like HIPAA and FERPA in the USA and GDPR in the EU are complex, and are refined over decades, but they have successfully achieved some degree of trustworthiness in specific domains.

Both overtrust and undertrust are potential perils of a new technology [5]. If we undertrust, and lack the ability to use surveillance data to meet existential threats, society could suffer grave damage. But overtrusting leaves us vulnerable to many kinds of exploitation. The viability of our society may depend on our ability to strike this balance.

For the issue of data, surveillance, and privacy, it is corporations and government agencies whose trustworthiness we must assess. As individual AI agents become more complex and sophisticated, and are called upon to make and defend more nuanced judgments, our

assessments of their trustworthiness will increasingly resemble our assessments of the trustworthiness of other humans. Drawing on reputation and observations of behavior, we will estimate factors of trustworthiness including ability, benevolence, and integrity [7].

The ETC framework focuses our attention on the need for trust, and therefore the critical need to ensure the trustworthiness of all sorts of agents: human, corporate, or AIs.

6 CONCLUSIONS

We now recognize that making decisions about industrial processes without considering their impact on the global environment has led to very serious problems with climate change. Similarly, making decisions about information processes and the interactions among humans, corporations, and AIs without considering their impact on trust and trustworthiness can lead to failures of cooperation and the weakening of society.

Our society is extremely complex, including many different issues which must be addressed systematically and comprehensively. For the issue of data, surveillance, and privacy, questions that help us define critical boundaries include (a) what individual members of society should be able to trust about their data, and (b) how the entities collecting our data can demonstrate and guarantee their trustworthiness. These questions about trust and trustworthiness must be asked about other issues in AI Ethics including bias and fairness, and the safety of intelligent systems.

7 ACKNOWLEDGMENTS

Thanks to Peter Railton, Michael Wellman, many students in my Ethics for AI and Robotics classes, and an anonymous reviewer.

8 REFERENCES

[1] Robert Axelrod. 1984. *The Evolution of Cooperation*. Basic Books, New York.

[2] Joseph Henrich. 2016. *The Secret of Our Success*. Princeton University Press.

[3] Benjamin Kuipers. 2012. An existing, ecologically-successful genus of collectively intelligent artificial creatures. *Proc. Collective Intelligence Conference*. arXiv:1204.4116.

[4] Benjamin Kuipers. 2022. Trust and cooperation. *Frontiers in Robotics and AI* 9:676767. doi:10.3389/frobt.2022.676767.

[5] J. D. Lee and K. A. See. 2004. Trust in automation: designing for appropriate reliance. *Human Factors* 46(1):50-80.

[6] Niklas Luhmann. 1979. Trust: A mechanism for the reduction of social complexity. In *Trust and Power: Two Works by Niklas Luhmann*. Wiley.

[7] R. C. Mayer, J. H. Davis and F. D. Schoorman. 1995. An integrative model of organizational trust. Academy of Management Review 20(3): 709-734.

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

[9] Michael Tomasello. 2019. *Becoming Human: A Theory of Ontogeny*. Harvard University Press.

[10] Robert Wright. 2000. *Nonzero: The Logic of Human Destiny*. Pantheon.

[11] Shoshana Zuboff. 2019. *The Age of Surveillance Capitalism*. PublicAffairs, New York.

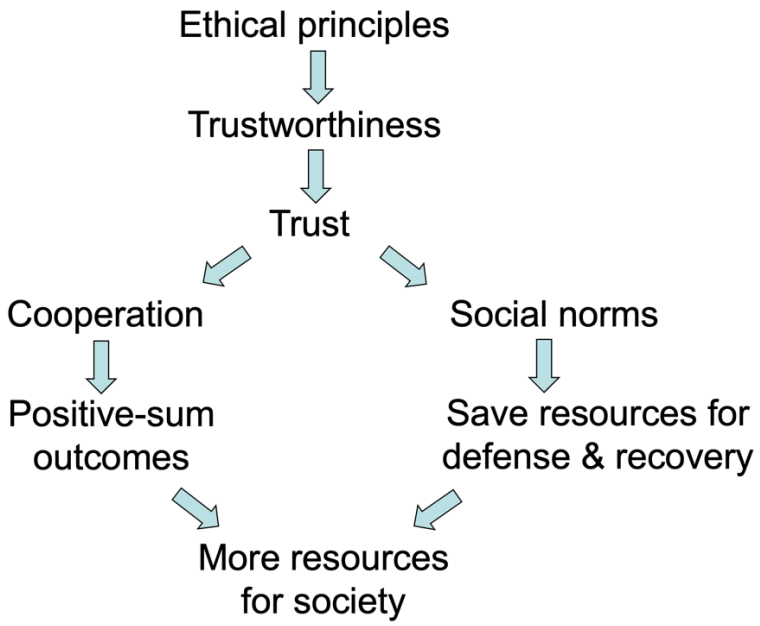

Figure 1: Trustworthiness and Trust are central concepts on a causal chain from Ethical principles to Resources for society. Cooperation involves known and trusted partners collaborating in a positive-sum activity. Social norms allow one to count on others who may not be known as individuals, avoiding costs for actively defending against, or recovering from, exploitation.