Perspectives on Ethics of AI: Computer Science*

Benjamin Kuipers†

August 14, 2019

Abstract

AI is a collection of computational methods for studying human knowledge, learning, and behavior, including by building agents able to know, learn, and behave. Ethics is a body of human knowledge, far from completely understood, that helps agents (humans today, but perhaps eventually robots and other AIs) decide how they and others should behave. The ethical issues raised by AI fall into two overlapping groups.

First, potential deployments of AI raise ethical questions about the impacts they may have on human well-being, just like other powerful tools or technologies such as nuclear power or genetic engineering.

Second, unlike other technologies, intelligent robots and other AIs have the potential to be considered as *members* of our society. Since they will make their *own* decisions about the actions they take, it is appropriate for humans to expect them to behave ethically. This requires AI research with the goal of understanding the structure, content, and purpose of ethical knowledge, well enough to implement ethics in artificial agents.

This chapter describes a computational view of the function of ethics in human society, and discusses its application to three diverse examples.

*Draft chapter for the *Oxford Handbook of Ethics of AI*, edited by Markus Dubber, Frank Pasquale, and Sunit Das, to appear, 2019.

†kuipers@umich.edu. Computer Science & Engineering, University of Michigan, Ann Arbor, Michigan 48109 USA

1 Why Is the Ethics of AI Important?

AI uses computational methods to study human knowledge, learning, and behavior, in part by building agents able to know, learn, and behave. Ethics is a body of human knowledge that helps agents (humans today, but perhaps eventually robots and other AIs) decide how they and others should behave. The ethical issues raised by AI fall into two overlapping groups.

First, like other powerful tools or technologies (e.g., genetic engineering or nuclear power), potential deployments of AI raise ethical questions about their impact on human well-being.

Second, unlike other technologies, intelligent robots (e.g., autonomous vehicles) and other AIs (e.g., high-speed trading systems) make their own decisions about the actions they take, and thus could be considered as *members* of our society. Humans should be able to expect them to behave ethically. This requires AI research with the goal of understanding the function, structure, and content of ethical knowledge well enough to implement ethics in artificial agents.

As the deployment of AI, machine learning, and intelligent robotics becomes increasingly widespread, these problems become increasingly urgent.

2 What is the Function of Ethics?

"At the heart of ethics are two questions: (1) What should I do?, and (2) What sort of person should I be?"[1] Ethics consists of principles for deciding how to act in various circumstances, reflecting what is right or wrong (or good or bad) to do in that situation.

It is clear that ethics (and hence what is considered right or wrong, or good or bad) changes significantly over historical time. Over similarly long historical time-scales, despite discouraging daily news reports, it appears that the societies of our world are becoming stronger, safer, healthier, wealthier, and more just and inclusive for their members.[2]

Two important sources of concepts help make sense of these changes. First, game theory contributes the abstraction of certain types of interactions among people as games[3], and behavioral economics shows that these games not only have winners and losers, but the overall impact on the players collectively can be described as positive-sum, zero-sum, or negative-sum.[4] Second, the theory of evolution, as applied to human and great ape cognition and sociality, shows how a way of life that depends on positive-sum cooperation among individuals is likely to

---

[1] Russ Shafer-Landau, editor. *Ethical Theory: An Anthology*. Wiley-Blackwell, second edition, 2013. p. xi.

[2] Robert Wright, *Nonzero: The Logic of Human Destiny*, Pantheon, 2000. Steven Pinker, *The Better Angels of Our Nature: Why Violence Has Declined*, Viking Adult, 2011. Steven Pinker, *Enlightenment Now: The Case for Reason, Science, Humanism, and Progress*, Viking, 2018.

[3] John von Neumann and Oskar Morgenstern. *Theory of Games and Economic Behavior*. Princeton University Press, 1953.

[4] Samuel Bowles, *The Moral Economy: Why Good Incentives are No Substitute for Good Citizens.* Yale University Press, 2016.

provide for its society greater fitness than less cooperative ways of life.[5] We can therefore think of the function of ethics as promoting the survival and thriving of the society by influencing the behavior of its individual members, summarized as:

> *Ethics is a set of beliefs that a society conveys to its individual members, to encourage them to engage in positive-sum interactions and to avoid negative-sum interactions.*

As a society prospers, survives, and thrives, its individual members benefit as well, so ethical behavior is "non-obvious self-interest" for the individual.

Philosophers would consider this to be a *rule consequentialist* position[6], but one where the relevant consequences are the survival and thriving of society, not the pleasures and pains of its individual members. It is *consequentialist* because actions are not evaluated according to whether they are intrinsically right or wrong (by some criterion), but according to their long-term good or bad consequences for the survival and thriving of society. This position is *rule* consequentialism because the unit that is evaluated is not the individual action decision, but the set of ethical principles (often rules) adopted by society.

Positive-sum and negative-sum interactions. Commerce and cooperation are paradigm positive-sum interactions. When one person voluntarily trades or sells something to someone else, each party receives something that they value more highly than what they gave. When cooperating on a project, partners contribute toward a common goal, and reap a benefit greater than either could achieve alone.

Theft and violence are examples of negative-sum interactions. The thief gains something from the theft, but the loss to the victim is typically greater than the gain to the thief. Violent conflict is the paradigm negative-sum interaction, since both parties may be worse off afterwards than before, possibly much worse off. (These are not cleanly separated cases. Violence in defense against external attack may be necessary to avoid a catastrophic outcome, and that defense itself is likely to be a cooperative project.)

Cooperation, trust, and social norms. Cooperative projects among individuals are a major source of positive-sum outcomes. However, cooperation requires vulnerability, and trust that the vulnerability will not be exploited.[7]

> *Trust is a psychological state comprising the intention to accept vulnerability based on positive expectations of the intentions or behavior of another.*[8]

---

[5] Michael Tomasello. *A Natural History of Human Morality*. Harvard University Press, 2016.

[6] Walter Sinnott-Armstrong. Consequentialism. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Winter 2015 edition, 2015.

[7] Michael Tomasello. *A Natural History of Human Morality*. Harvard University Press, 2016.

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

As intelligent robots or large corporations increasingly act as autonomous goal-seeking agents and therefore as members of our society, then they, too, need to be subject to the requirements of ethics, and need to demonstrate that they can trust and be trustworthy.

Successful cooperation demonstrates the trustworthiness of the partners and produces more trust while exploitation reduces trust. By trusting each other enough to pool their resources and efforts, individuals working together can often achieve much more than the sum of their individual efforts working separately. Large cooperative projects, from raising a barn, to digging a canal, to creating an Interstate Highway System, produce large benefits for everyone. But if I spend a day helping raise your barn, I trust that in due time, you will spend a day helping to raise mine. And if taxes help pay for New York's Erie Canal or the Pennsylvania Turnpike, I trust that, in due time, taxes will also pay for the Panama Canal linking the East and West Coasts, and the St. Lawrence Seaway providing access to the Great Lakes. Some of the states in the USA emphasize this with the name "Commonwealth", meaning that shared resources provide shared prosperity.

Social norms are behavioral regularities that we as individual members of society can generally count on when planning our activities. By trusting these (near) invariants, many aspects of our lives become simpler, more efficient, and less risky and uncertain. Maintaining a social norm is a kind of cooperative project without specified partners. I accept certain minor sacrifices in return for similar behaviors by (almost) everyone else, providing a (near) invariant that we all can rely on.

For example, when having lunch at a cafe, condiments are freely available for my convenience, but I know not to pocket the extras, so they will continue to be available. Likewise, I trust that a simple painted stripe in the middle of a road I am driving on securely separates me from drivers going in the opposite direction, so I accept the minor sacrifice of not crossing that stripe even when my side is congested.

Like explicit cooperative projects, social norms provide positive-sum results for society, saving resources that would otherwise go toward protection and recovery, making us individually and collectively better off. Each requires *trust*: acceptance of vulnerability to the other partners, along with confidence that few others will exploit that vulnerability, even for individual gain.

I use the term "social norm" inclusively, to cover regularities ranging from laws and moral imperatives to non-moral social conventions. Philosophers make many different distinctions among types and origins of social norms. By taking a design stance toward ethical systems for influencing the behavior of intelligent agents, human and non-human, in our society, I emphasize the common functional goal of encouraging positive-sum, and discouraging negative-sum, interactions.

Representing ethical knowledge. I have described ethics as "*a set of beliefs that a society conveys to its individual members*", and have stated that those beliefs are evaluated according to "*their long-term good or bad consequences for the survival and thriving of the society*." Since the result of this evaluation depends on many complex factors and evolves over decades and centuries, it is not very useful to individuals in deciding how to act.

To make practical decisions, individual humans need concise and understandable ethical principles. For these principles to be useful for the long-term survival of the society, they must also be explainable and teachable to individuals entering the society, such as children and immigrants. If intelligent non-human agents such as robots and corporations are to apply ethical principles to their own behavior, these principles must be capable of being learned or programmed.

The field of philosophical ethics has, over the centuries, created a number of concise frameworks for ethical knowledge, built around concepts such as virtues, duties, utilities, contracts, etc.[9] While it is tempting to regard these as competing alternatives, it is generally recognized that they are pieces of a more complicated, incompletely understood, puzzle (cf. John Godfrey Saxe's 1873 children's poem, *The Blind Men and the Elephant*).

The many fields of applied ethics (e.g., biomedical ethics[10]) appeal to all of these conceptual frameworks, starting with specific ethical questions and searching for clear, practical answers. Depending on the details of the case in question, clarity may come from one or another of the ethical frameworks, while others provide ambiguous or unacceptable results.

It may be possible to express several of these conceptual frameworks in a single knowledge representation based on *cases*, $\langle S,A,S',v \rangle$, where S and S' represent previous and resulting situations, A describes an action, and v is an evaluation.[11] The representation can describe the situations and action at different levels of detail, ranging from rich descriptions of experienced events, to highly schematic general patterns.

Ethics Research in the AI Community. A number of AI and robotics researchers explicitly address the problem of ethics for AI and robotics.[12] For example, Ron Arkin proposed that an autonomous system controlling a lethal weapon could be equipped with an "ethical governor" based on the Laws of War and Rules of Engagement, with the authority to override an attempt to deploy lethal force.[13] Human emotional reactions can lead to errors and even war crimes. Arkin claims that, by taking the human out of the loop, targeting can be more precise and lawful, making war more humane. Many others are more skeptical about the impact of lethal autonomous weapon systems.

Utilitarianism has been attractive in the AI community because it factors ethical decisions into (a) defining a utility function that represents preferences over states of the world, and (b) applying an optimization algorithm to identify the action (or rule) that maximizes expected

[9] Russ Shafer-Landau, editor. *Ethical Theory: An Anthology*. Wiley-Blackwell, second edition, 2013.

[10] T. L. Beauchamp and J. F. Childress. *Principles of Biomedical Ethics*. Oxford University Press, sixth edition, 2009.

[11] B. Kuipers. How can we trust a robot? *Communications of the ACM*, **61**(3):86–95, 2018.

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

[13] Ronald C. Arkin. *Governing Lethal Behavior in Autonomous Robots*. CRC Press, 2009.

utility. While philosophical utilitarianism aggregates utility over everyone in the society[14], in game theory each individual player optimizes his/her own utility.[15]

A motivating problem is that there are many cases (e.g., the Prisoners' Dilemma, the Public Goods Game, the Tragedy of the Commons. etc.) where the "rational" solution according to game theory (the Nash equilibrium) results in poor outcomes for every player and a negative-sum result for the society. And in fact, humans playing these games tend to avoid the Nash equilibrium and get better outcomes.[16]

Much effort has gone into formulating utility functions for individual decision-making that lead to improved outcomes for everyone in society, often in the context of repeated games drawing from the same population of players. Vincent Conitzer and colleagues[17] show how a player can communicate its intention to behave in a trustworthy way by making a "sub-optimal" move. The other player is meant to understand this as an offer to cooperate, and feel obligated to reciprocate. Stuart Russell and others have posed the problem of *value alignment*,[18] as defining utility functions that lead to decisions similar to those that humans make. *Cooperative inverse reinforcement learning*[19] has been proposed as a solution to the value alignment problem where the robot tries to maximize the *human's* utility function, while recognizing that it has only incomplete knowledge of that utility function. This is intended to prevent a robot, however powerful, from optimizing a poorly chosen utility function in a way that causes a catastrophe according to human utilities.[20]

3 Human and Non-Human Members of Society

Traditionally, a society's members are the individual human beings who participate in the society by interacting with each other and making decisions about what actions to perform.

In recent years, progress in artificial intelligence, robotics, and machine learning has raised the prospect of intelligent non-human robots participating as members of our society. Autonomous vehicles must be trusted to behave safely and ethically in both routine traffic and emergency situations.[21] Other AIs that are not physically embodied, such as high-speed trading systems or social networks, should also behave safely and ethically.[22] Large-scale institutions can also be

---

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

[15] John von Neumann and Oskar Morgenstern. *Theory of Games and Economic Behavior*. Princeton University Press, 1953.

[16] J. K. Goeree and C. A. Holt, Ten little treasures of game theory and ten intuitive contradictions, *The American Economic Review* **91**(5): 1402-1422, 2001. J. R. Wright and K. Leyton-Brown, Predicting human behavior in unrepeated simultaneous-move games, *Games and Economic Behavior* **106**: 16-37, 2017.

[17] J. Letchford, V. Conitzer, and K. Jain. An "ethical" game-theoretic solution concept for two-player perfect-information games. In *Int. Workshop on Internet and Network Economics (WINE)*, 2008.

[18] S. Russell, D. Dewey & M. Tegmark. Research priorities for robust and beneficial artificial intelligence. *AI Magazine* **36**(4): 105-114, Winter 2015.

[19] D. Hadfield-Menell, A. Dragan, P. Abbeel, and S. Russell. Cooperative inverse reinforcement learning. In *Advances in Neural Information Processing Systems (NIPS)*, 2016.

[20] Nick Bostrom. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, 2014.

[21] Patrick Lin. The ethics of autonomous cars. *The Atlantic Monthly*, 8 October 2013.

[22] M. P. Wellman and U. Rajan. Ethical issues for autonomous trading agents. *Minds & Machines* **27**: 609-624, 2017.

considered as intelligent entities: for-profit and non-profit corporations, governments, churches, unions, and other corporate entities.[23]

For all of these entities participating in society, the function of ethics is the same – to encourage positive-sum interactions and discourage negative-sum ones, supporting the survival and thriving of society as a whole. Likewise, the same means help to accomplish this function – supporting trust in relevant social norms, and for each entity to demonstrate that it is trustworthy.

4 Method: Analyzing Specific Cases of Trust and Ethics

There are many different domains of behavior, with different social norms and ethical principles available for trust. Furthermore, as noted before, social norms and ethical principles change over historical time. Our goal here cannot be to provide universal answers about how humans and non-human agents in society should behave. Rather, our goal must be to provide a framework for asking useful questions.

In the following sections, I discuss three quite different cases of ethical decision-making that are relevant to societies including both human and non-human agents. Autonomous vehicles are individual, embodied robots that make decisions about driving, some with ethical implications. Social networks are disembodied intelligent systems that mediate interactions among people, but that also collect large amounts of information, often disregarding individual privacy concerns. Corporations, to which we have entrusted much of the wealth in our economy, can also be viewed as intelligent agents, whose behavior should be governed by ethics.

In each of these examples, we ask what social norms people would want to trust. The ethical principles that a society adopts and encourages its individual members to follow determines the social norms that individuals in that society should be able to trust. We will consider how those social norms might be expressed.

5 Example 1: Trust and Ethics for Autonomous Vehicles

Vast sums are being invested to develop autonomous vehicles (AVs), which are intelligent robots intended to share the roads with ordinary human-driven vehicles as well as with pedestrians. These robots take passengers or cargo to their destinations, or simply bring the AV where it is next needed. The critical technological requirement is for the robot's perception to provide sufficient situational awareness, and for it to make the right decisions, to keep itself and humans safe.

To accept AVs on our roads, humans will need to trust their behavior. Inspired by Isaac Asimov's First Law of Robotics[24], "*A robot may not injure a human being or, through inaction, allow a human being to come to harm*", we might start by proposing the following social norm:

---

[23] B. Kuipers. An existing, ecologically-successful genus of collectively intelligent artificial creatures. In *Collective Intelligence*, 2012. arXiv:1204.4116.

[24] Isaac Asimov. *I, Robot*. Grosset & Dunlap, 1952.

(SN-0) A robot (or AI or AV) will never harm a human being.

This is overly sweeping, to the point of impossibility, even without the clause about not failing to prevent harm through inaction. However, if we distinguish between deliberate and accidental harm, we can formulate a pair of more plausible social norms:

(SN-1) A robot will never deliberately *harm a human being.*

(SN-2) In a given situation, a robot will be no more likely than a skilled and alert human to accidentally *harm a human being.*

Achieving these two social norms will require technical solutions to difficult problems in perception, situational awareness, planning, and acting, but they do not set the impossible goal of guaranteeing that fatal accidents can never occur. We still need a carefully stated social norm describing when action is required to prevent harm that would otherwise happen.

The Deadly Dilemma. A concerned philosopher, inspired by the famous "Trolley Problem"[25], might ask what the AV should do if it is suddenly confronted with a "Deadly Dilemma", where it cannot avoid colliding with one of two groups of humans, and must decide which group to deliberately kill. Either choice in this dilemma clearly violates the social norm (*SN-1*), and therefore undermines trust in AVs by members of society.

While the original Trolley Problem is a useful thought experiment that philosophers use to explore the relationships between human moral intuitions and the predictions of different philosophical theories, it is not a useful guide for the design of embodied robots in the physical world. To design an ethical robot (such as an AV), we must reject the narrow framing of the Trolley Problem, and formulate an additional social norm.

When humans experience a bad outcome, they often engage in *counterfactual thinking*, searching by mental simulation for a previous ("upstream") action that would have avoided the bad outcome.[26] For a unique event, counterfactual thinking is futile and can lead to depression, but for recurring types of events, it can produce valuable insights, "practical wisdom"[27], that leads to better outcomes in the future. A situation like the Deadly Dilemma, with no good outcomes, should trigger counterfactual thinking, so the driver learns that a previously-unremarkable situation like entering a narrow street requires driving much slower, to preserve the option of a safe emergency stop. By learning from counterfactuals, the attentive agent accumulates a store of practical wisdom that makes safe and ethical behavior much easier.

(SN-3) A robot must learn to anticipate and avoid Deadly Dilemmas.

---

[25] Judith Jarvis Thomson. The trolley problem. *Yale Law Journal*, 94(6):1395–1415, 1985.

[26] Neal Roese and Kai Epstude. The functional theory of counterfactual thinking: New evidence, new challenges, new insights. *Advances in Experimental Social Psychology* **56**: 1—79, 2017.

[27] Aristotle. Nicomachean Ethics. Translated by Terence Irwin, Hackett, Second edition, 1999.

The concerned philosopher responds, "Yes, this scenario is unlikely, but what if it *does* happen?"

Perception in the physical world is imperfect, so neither humans nor robots can perceive an emergency situation well enough to be certain that it presents a Deadly Dilemma between exactly two alternatives. There is a probability distribution over a continuous space of similar scenarios, some of which involve fatalities, while many others are "Near Misses". A Near Miss is far more likely than a true Deadly Dilemma.

$$p(NearMiss \mid Observation) \gg p(DeadlyDilemma \mid Observation).$$

The best response when suddenly confronted by this situation is immediate emergency braking along with steering to minimize risk of injuries. This response satisfies the two social norms: (*SN-1*) the robot does not deliberately target any human, even to save others; and (*SN-2*) its probability of injuring a human is no greater than for a skilled and attentive human driver, faced with the same situation. Even in the rare case that there is a fatality, the AV has acted reasonably and ethically when confronted by a bad situation.

Aristotle tells us that virtue is a skill that improves with experience, like carpentry. The novice may be presented with a situation that appears to be a Deadly Dilemma. The expert has more experience, more practical wisdom, and acts earlier so the Deadly Dilemma can be avoided.

Ethical Principles to Encourage Trust. The social norm (*SN-1*) above translates naturally into an easily stated ethical duty: *Never deliberately harm a human being*. To the extent that a robot visibly follows this rule, it becomes more *trustworthy*, and is increasingly trusted to follow the rule in the future.

The second social norm (*SN-2*) sets a bar for competence. The capabilities of human drivers and AVs can be tested and compared. Young humans are subject to age, time, and situation constraints on driving, until they accumulate enough experience and practical wisdom to become trustworthy drivers. Likewise, elderly human drivers face ethical requirements to restrict or give up their own driving according to their abilities as observed by themselves or others.

The third social norm (*SN-3*) requires a continual effort to anticipate potential Deadly Dilemmas via counterfactual thinking, learning to recognize the upstream decision point and the choice that avoids the Dilemma.

As engineered devices, AVs can be designed with mechanisms for self-monitoring and self-evaluation, to determine in real time whether they are able to drive safely in the current situation. The details of such mechanisms may not have concise descriptions in natural language, but their overall effect would correspond to an ethical duty such as: *When it is not safe to drive, stop safely and ask for assistance*.

Many other circumstances can arise when an AV shares our roads with human drivers and pedestrians. For example, if an AV is stopped at a cross-walk, how can a human pedestrian trust it enough to walk in front of it? This requires adequate situational awareness by the AV, and also the ability to communicate its trustworthiness to the human pedestrian. Both of these problems

may have technical solutions, but even a restricted domain like driving includes a very large number of these problems.

Over the centuries, human societies have accumulated huge numbers of situation-specific social norms to trust, along with ways for agents to signal their trustworthiness, and both society and the lives of its individual members, have improved as a result.[28]

6 Example 2: Individual User Models

People are complex, and so is our world. We have incomplete understanding of our world, of each other, and of ourselves. We love to communicate with each other, and we depend on that communication, including the feedback we get from others, to create, develop, correct, and refine our understanding of reality.

Human experience with intelligent agents is almost entirely with other humans, where different capabilities are highly correlated. We humans are prone to anthropomorphize non-human, and even inanimate, elements of our environment where we can attribute agency.[29] This can easily lead to assuming that robots and other AIs are more human-like and more capable than they actually are.[30] Generalizations that are useful with other humans are unreliable with robots and other AIs, possibly leading to excessive trust, unexpected catastrophes, and other ethical problems.

We use search engines (like Google) to find what other people have written or created. We use social networks (like Facebook) to communicate with each other about what we are doing, and to learn about what they are doing. We understand that these services cost money, and they have to be paid for somehow. We have long accepted that advertisements help pay for newspapers, magazines, and television. Modern data mining methods, using new machine learning algorithms, vast quantities of data, and abundant computing resources, have made it increasingly feasible to build detailed models of individual users. Without a deep understanding of what these websites do and how they do it, we extend our acceptance to the creation of individual user models that can be sold to advertisers to improve the targeting of their advertisements. Many users consider it worthwhile to trade some of their privacy for "free" search and social network services, paid for by advertising that is better matched to their own personal interests.

This use of individual user models could be seen as an ethically acceptable bargain, satisfying a social norm of the form:

> *(SN-4) I understand that Internet companies earn money by creating models of me and my interests from the information I knowingly and voluntarily provide, and*

---

[28] Steven Pinker. *Enlightenment Now: The Case for Reason, Science, Humanism, and Progress*. Viking, 2018.

[29] N. Epley, A. Waytz & J. T. Cacioppo. On seeing human: A three-factor theory of anthropomorphism. *Psychological Review* **114**(4): 864-886, 2007.

[30] P. Robinette, R. Allen, W. Li, A. M. Howard & A. R. Wagner. Overtrust of robots in emergency evacuation scenarios. *ACM/IEEE Int. Conf. Human-Robot Interaction* (*HRI*), pages 101-108, 2016.

selling access to those models to advertisers. I trust that the advertisers will use these models to serve me with ads that better match my personal interests.

The individual users of Google's search engine or Facebook's social network (or many other useful apps) are the sources of data from which the models are built. We would like to trust social norms such as:

(SN-5) Except for clearly marked advertisements, the results from a search are the AI's best attempt to understand what I want, and retrieve answers to my questions and access to desired Internet sites.

(SN-6) Except for clearly marked advertisements, a social network presents me with a reasonably unbiased sample of the posts created by people linked to me in the network. They receive my posts via a similarly unbiased sampling algorithm.

In many cases, we *do* trust these social norms. In the real world, the evidence suggests that this trust is *not* justified.[31] Specifically, Google, Facebook, and other major Internet companies collect and aggregate far more behavioral information about individual users than we "knowingly and voluntarily provide" (violating *SN-4*). Furthermore, the results they return are designed to influence our future behavior beyond our shopping choices. We are naive to trust that these systems are unbiased and non-manipulative (i.e., they violate *SN-5* and *SN-6*).

The perils of correct individual models. Individual users typically don't understand the breadth of data that these model-builders can draw on. Internet companies can collect information not only from direct interaction with their own interfaces, but also from interactions with other sites, from "cookies" left behind with tracking information, and from many other observation channels.

Most Internet users have had experiences like the following, or worse. Once I did a Google search in one browser for a style of dining-room chair I found attractive. Shortly afterward Facebook, running in a different browser, began serving me ads for that style of chair. This felt creepy, like "telepathic" surveillance of my personal interests and activities. My dining-room-chair preferences are not particularly sensitive information, but who knows what other kinds of surveillance they are doing?

In normal human communication, many of the things we communicate via speech, text, or email are *ephemera* – temporary statements that may be context-dependent, poorly thought out or poorly stated, intended to be refined or discarded in the course of the conversation. And they are communicated with different individuals, who we trust are not conspiring to assemble comprehensive models of our preferences, beliefs, personalities, and activities.

(SN-7) I trust that small pieces of information, shared with different agents, will not be aggregated and correlated to create an inappropriately invasive model of me as an individual, violating my privacy.

---

[31] Shoshana Zuboff. *The Age of Surveillance Capitalism*. Public Affairs, New York, 2019.

This is, of course, exactly what major Internet companies like Facebook and Google do with their machine learning algorithms and access to vast streams of data.[32] Even if the models they create are correct, their predictions are likely to invade my privacy.

I have a right to keep actions and beliefs to myself, if I don't want to share them with others. One anecdote tells of a young man who bought a diamond ring online, intending to surprise his girlfriend with a marriage proposal, but the merchant sent email to all his Facebook friends, congratulating him on his engagement. This was a minor annoyance, but similarly inferring and broadcasting the political actions or opinions of a person living in a repressive state could be life-threatening.

Insurance companies are among the many companies taking advantage of the Internet of Things (IoT) to gather surveillance information about individual behavior. Both auto and health insurance companies can increasingly monitor compliance with various constraints, punishing violations with increasing premiums, insurance cancellation, or even by disabling the car.

"Legals", including End-User License Agreements (EULAs), Privacy Policies, and Terms of Service, are the long, dense, legal agreements that most of us click through without reading, in order to gain access to software, "free" or otherwise. These agreements authorize the company providing the software to collect our data and to share it with, or sell it to, other companies, typically without meaningful constraint. "Legals" are designed to discourage users from reading them, and they allow the companies to claim that users voluntarily "opt in" to these data sharing conditions.

An analysis of the legal agreements associated with the Nest "smart" thermostat[33] found (sect. 4) that if a UK-based customer wants "a comprehensive picture of the rights, obligations and responsibilities of the various parties in the supply chain, he has to read at least 13 legal items." Worse, those link to additional contractual agreements from partners, affiliates, manufacturers of interoperable products, and others. Following these links, "If you add to Nest legals those of the connected devices, apps and appliances, the result is that for what appears to be a single product, a thousand contracts may apply!"

During the 2016 U.S. Presidential election campaign, the company Cambridge Analytica used Facebook data to build models identifying people who were vulnerable to conspiracy theories, and targeted them for ads motivating them to turn out and vote for a particular candidate.[34] Even if most people are correctly confident in their own resistance to such ads, *some* people can be manipulated by unscrupulous advertisers, and their votes may affect the outcome for everyone.

Internet companies sometimes argue that their user modeling technologies are morally neutral, and that it is only the application of those models by companies like Cambridge Analytica that

---

[32] Shoshana Zuboff. *The Age of Surveillance Capitalism*. Public Affairs, New York, 2019.

[33] G. Noto La Diega and I. Walden. Contracting for the 'Internet of Things': looking into the Nest. *European Journal of Law and Technology*, 7(2), 2016.

[34] Nicholas Confessore. Cambridge Analytica and Facebook: The Scandal and the Fallout So Far. *The New York Times*, 4 April 2018. Retrieved 5-17-19 from https://www.nytimes.com/2018/04/04/us/politics/cambridge-analytica-scandal-fallout.html.

raises ethical problems.[35] However, when Google and Facebook sell tools and access to data that makes it easy and profitable for others to violate our privacy, or manipulate the institutions of our society, surely they are not absolved from ethical responsibility!

The perils of incorrect individual models. Incorrect user models can cause problems ranging from the trivial (display of irrelevant ads) to life-transforming (denial of probation or bail). A learning system can pick up biases from its training data, possibly from unconscious bias in how it is assembled, possibly because of the impact of historical bias on the phenomena being measured.

Sometimes, a model is incorrect because the designers of the system made grossly incorrect assumptions. Starting in October 2013, the Michigan Integrated Data Automated System (MiDAS) automatically evaluated claims for unemployment insurance.[36] Any information discrepancy between the applicant and the employer was treated as evidence of fraud by the applicant. A letter was generated and sent to the applicant's last known address. If not returned within 10 days, the applicant was considered guilty, and the algorithm immediately imposed major financial penalties, with no human review, causing great hardship. A review of 22,427 charges filed between 2013 and 2015 revealed a 93% error rate!

It is now widely known that automated face detection and face recognition systems often have significantly higher error rate for faces with darker skin.[37] This can happen even though the algorithm learns correctly from the training examples, because the set of training examples does not adequately reflect the diversity of the population. Similar problems occur in medical diagnosis: male and female patients having a heart attack exhibit significantly different symptoms. In decades past, most data for the study of heart attacks came from male patients, leading to frequent misdiagnosis for female patients.[38] Efforts are under way to redress these data imbalances, but much remains to be done.

In other cases, the training set could perfectly reflect human behavior, but that behavior includes the effects of existing biases. Finding ways to train a complex machine learning system, while avoiding biases that may be embedded in the training data, is a difficult open problem.[39]

Membership in a particular minority group may be genuinely statistically correlated, in our society, with some characteristic of interest. But a fundamental principle in our society is that individuals should be judged as individuals, without bias from membership in a particular

---

[35] *"Once the rockets are up, Who cares where they come down. That's not my department!" Says Werner von Braun.* [Tom Lehrer, *That Was The Year That Was*, 1963]

[36] Ryan Felton. Michigan unemployment agency made 20,000 false fraud accusations – report. *The Guardian*, 18 December 2016. Retrieved 5-17-19 from https://www.theguardian.com/us-news/2016/dec/18/michigan-unemployment-agency-fraud-accusations.

[37] B. Wilson, J. Hoffman, and J. Morgenstern. Predictive inequity in object detection. Technical Report arXiv:1902.11097, ArXiv, 21 Feb 2019.

[38] T. A. Beery. Gender bias in the diagnosis and treatment of coronary artery disease. *Heart & Lung* **24**(6): 427-435, 1995.

[39] Cathy O'Neil. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Penguin Random House, 2016.

minority group.[40] It remains difficult to translate this societal ideal into inference methods for data analysis.

Conclusion. We live with intelligent tools and systems that are designed to satisfy our human needs and desires and provide their corporate owners with continuing streams of data about ourselves. Google (for access to information) and Facebook (for social communication) are only the beginning. They are designed to be addictive, so we keep interacting with them. They can learn a great deal about us, which makes them more valuable as tools for us, and also more valuable commercially, for selling individual user models to advertisers and others.

We trust that these intelligent systems follow social norms that we have learned from our experience interacting with other humans and with human-scale organizations. We have only begun to grapple with the impact of the vastly greater scale of the information involved, in terms of the number of people, events, and actions under surveillance; the microscopic detail of the information that can be collected, aggregated, and analyzed; the mass of training data that can be used to create predictive models of each individual; and the ways those predictions can be used for economic and political ends.

A homely example illustrates the impact of scale. If you are hiking alone, it is no problem to pee in the woods. The ongoing physical, biological, and social processes in the woods can handle that tiny load. But a city of 100,000 people is legitimately required by state and federal regulations to build an elaborate infrastructure to protect the physical, biological, and social environment, including water and sewage systems and a sewage treatment plant.

We are accustomed to broadcast ads that help support newspapers, magazines, and television. We accept political campaigns sending volunteers to knock on the doors of their supporters, to get out the vote on election day. We understand that every interaction reveals a little bit about ourselves. Once upon a time, the human scale and human limitations of these interactions provided implicit protection from many potential problems. But those times, and the scale of data collection, have changed.

We as a society don't grasp the implications of the massive change in scale – size, scope, detail, pervasiveness – that the development and deployment of surveillance capitalism brings.[41] We don't yet have a clear understanding of what we need to protect, how different kinds of costs and benefits trade off in this space, and what regulations we need. [42]

Large complex systems require large complex regulations. Those regulations necessarily evolve over time as we debug and refine them, and as society's understanding of its needs changes. Our society does have relevant large-scale experience with dissemination and protection of large amounts of data, including the FDA (US Food and Drug Administration, 1906) that ensures the safety and quality of food, drugs, and many other products; the SEC (US Securities and

---

[40] "*I have a dream that my four little children will one day live in a nation where they will not be judged by the color of their skin, but by the content of their character.*" Martin Luther King, Jr. [28 August 1963, March on Washington]

[41] Cathy O'Neil. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Penguin Random House, 2016.

[42] Shoshana Zuboff. *The Age of Surveillance Capitalism*. Public Affairs, New York, 2019.

Exchange Commission, 1934) that regulates the nation's securities industry; FERPA (Federal Educational Rights and Privacy Act, 1974) that protects student educational records; HIPAA (Health Insurance Portability and Accountability Act, 1996) that protects personal medical information; and GDPR (EU General Data Protection Regulation, 2016) that protects data and privacy within the European Union.

A *fiduciary* is a person or organization that acts as a trustee for one or more beneficiaries, for example the asset manager of a pension fund or the trust department of a bank. A fiduciary has the duty to avoid any kind of conflict of interest, and to act solely in the beneficiary's interest. Fiduciary relationships are most common in financial domains, but the fiduciary concept also applies in other spheres.

Should companies like Facebook and Google, that collect and aggregate large amounts of personal data, have a fiduciary duty toward their individual users, requiring them to handle that data in the users' interest? The users' interest can certainly include personalized advertising that more closely aligns with individual preferences, and personalized recommendations of books, music, and other products based on previous choices. As long as the beneficiary is not exploited, it is not necessarily a conflict with its fiduciary duty for the company that collects and analyzes the data to profit from its efforts.

On the other hand, some current practices would violate those fiduciary duties. Click-through "agreements" that are designed to obtain legal "opt-in" permission while discouraging meaningful consideration of their conditions are clearly not in the user's interest. Similarly, meaningless "permission" for data sharing with other organizations, requiring the individual user to find and check the privacy policies of those other organizations, would violate the fiduciary duties. Where data sharing is needed for subcontracting some of the work, or for a business partnership, the original company must be responsible for ensuring that the partner provides protections at least as strong as the original company.

Like the GDPR in the EU, the details of such a fiduciary duty would be negotiated as legislation is designed, and then refined in the courts. The important point is to create a social norm that each individual can trust, along with meaningful enforcement mechanisms:

> *[SN-8] An organization that systematically collects, aggregates, and analyzes personal data about me is subject to a fiduciary duty to use that data in my best interest.*

7 Example 3: Sharing the Wealth

Fairness is important to adult humans, to children including young infants[43], and even to some species of non-human primates.[44] One way to study fairness in the laboratory is the *Ultimatum game*:[45]

> *The Ultimatum Game has two participants, A and B. A is given a sum of money, say $100. He may split this with B as he wishes. B may accept the offer from A, or he may reject it, in which case neither participant gets anything.*

The Nash equilibrium solution from game theory is clear: *A* makes the minimal offer to *B*, say $1, which *B* accepts, since $1 is better than nothing. The behavior of human participants is quite different: *A* tends to offer $40-50, and *B* tends to reject offers less than about $30. Often, *B* is willing to accept a substantial loss to punish *A* for making an unfair offer.

The total productivity of American society, and hence its total wealth, have been increasing steadily since the end of World War II. Much of that wealth is controlled by corporations, which historically responded to the needs of various *stakeholders*, including shareholders, workers, customers, suppliers, and neighbors. As the wealth of our society grew, the prosperity of the typical worker the United States increased at about the same rate for several decades (Figure 1(left)). People trusted that the economy would be *fair*:

> *(SN-9) Those who contribute to the success of a collective effort, will share in the benefits.*

Starting in the 1960s, Milton Friedman[46] and others argued that a corporation is purely a mechanism for maximizing wealth for its shareholders. The corporation and its human managers have responsibilities, but only to the *shareholders*, and not to other stakeholders such as workers, customers, suppliers, and neighbors, except as their responses might affect shareholder value. This change in the perceived ethical responsibilities of corporations has been widely accepted, especially by the business community.

The overall steady growth in wealth has continued, but starting around 1980, income gain became almost flat for the lower half of the economy. This has led to a dramatic increase in inequality among individuals, with most gains going to the top 1% of the population, and even more dramatically to the top .01% (Figure 1(right)).

---

[43] S. Sloan, R. Baillargeon & D. Premack. Do infants have a sense of fairness? *Psychological Science* **23**(2): 196-204, 2012.
[44] S. F. Brosnan & F. B. M. de Waal. Monkeys reject unequal pay. *Nature* **425**: 297-299, 2003.
[45] M. A. Nowak, K. M. Page & K. Sigmund. Fairness versus reason in the Ultimatum Game. *Science* **289**: 1773-1775, 2000
[46] Milton Friedman. The social responsibility of business is to increase its profits. *The New York Times Magazine*, 13 September 1970.

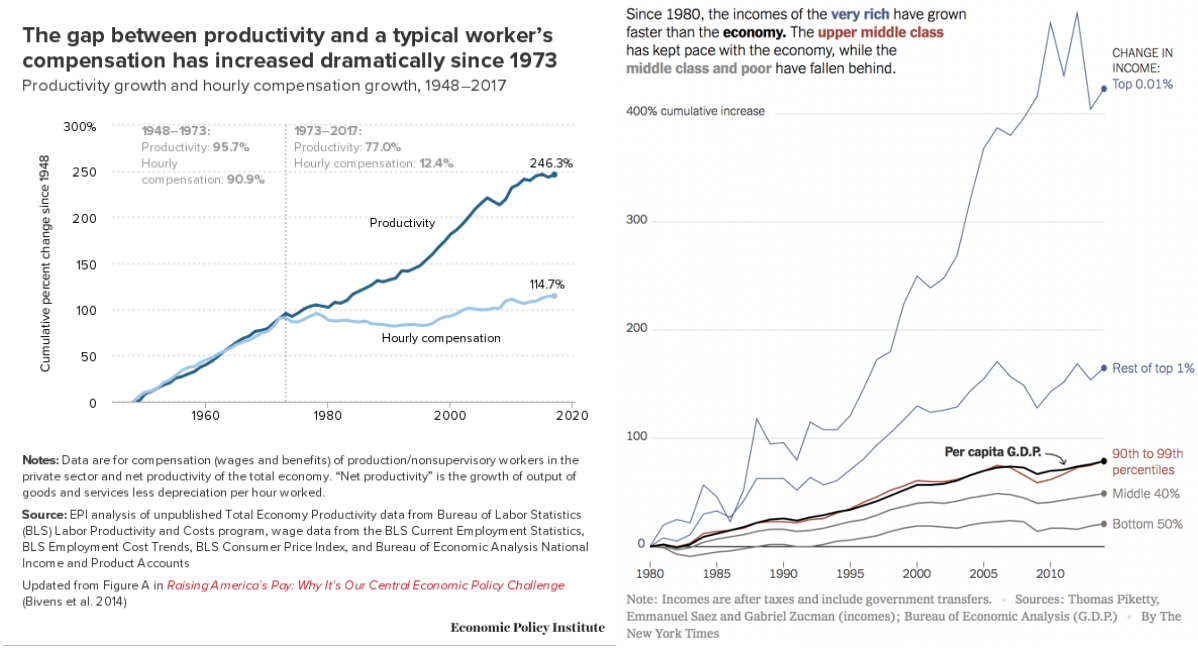

Figure 1: (left) Productivity and therefore national wealth have increased steadily since the late 1940s, but typical worker compensation leveled off in the mid-1970s.[47] (right) After about 1980, the incomes of the upper-middle class (90-99%) tracked the increase in per capita GDP, with the upper 1% increasing above that rate, and the lower 90% falling behind.[48]

The economics and the politics of our society have changed from offering opportunity for all, to one where the rich get ever richer, and the poor lose what little they had, even hope for the future and for their descendants. As these trends continue, more people become convinced that the social norm *SN-9* has been broadly violated, and their share in the growing wealth of society has been taken from them. Hopelessness, anger, and lack of trust continue to grow, to the point where, as in the story of Samson in the Old Testament (*Judges 16:29-30*), they are prepared to pull down the pillars of society to destroy their tormenters as well as themselves. We see this in a growing polarization of our society.[49]

Accumulating anger and resentment amplify fears of a future in which AI and robotics increasingly take over the jobs that people depend on for their livelihoods.

Can we create new jobs? It is often said that, in previous periods of rapid technological change, more jobs were created than were lost. There could be significant dislocation, perhaps for decades, since the people who had lost jobs were not necessarily qualified for the new jobs, but in the long run, plenty of new jobs were created. Others respond that previous technological advances provided automated substitutes for human and animal strength and mechanical skill,

[47] Reproduced with permission from "The Productivity-Pay Gap", Economic Policy Institute, August 2018.

[48] From The New York Times, 2-24-2019. © 2019 The New York Times Company. All rights reserved. Used under license.

[49] Susan McWilliams. This political theorist predicted the rise of Trumpism. His name was Hunter S. Thompson. *The Nation*, December 15 2016. https://www.thenation.com/article/this-political-theorist-predicted-the-rise-of-trumpism-his-name-was-hunter-s-thompson/.

but current AI-driven advances provides substitutes for intelligence, and it is not obvious where we go from here.

However, if we look carefully for a scenario where plenty of new jobs are created, the outlines of a possible solution seem to appear. This exercise identifies three important "pieces of the puzzle", and focuses our attention on the question of how they can fit together.

First, as we have seen (Figure 1(left)), productivity and wealth in our society are increasing steadily, and this increase seems likely to continue. The driving force behind automation is the prospect that corporations can become ever more profitable by using AI and robotics to automate increasing aspects of production costs.

Second, it is clear that people need meaningful work, not just guaranteed income.[50] It is important for people as individuals to be engaged in cooperative efforts that they consider meaningful and important, and that benefit more than just themselves – their family, their community, their country, or the society as a whole. Society benefits from the positive-sum nature of cooperative effort, and also from its individual members being capable of skilled, disciplined, responsible work toward shared goals.[51]

Third, there are plenty of jobs requiring skills, commitment, and effort, and that substantially benefit society. The problem is that, in our current economy, many of these jobs are not net generators of profit for an employer, so without subsidies, such jobs will not be created and filled.

One example of such a job is stay-at-home parent of young children. Such a job has substantial benefits for the children, for the family, and for the local community. When performed by a parent who wants to do this work, it cultivates skills, commitment, and effort, and can be extremely satisfying. However, it is not a profit center for our economy. It is typically unpaid, with a family unit supporting one person to do this work with little or no external financial support.

Another example is a job as a professional care-giver for children or the elderly. This job is essential where care for dependents is necessary, but family members must work for pay. Jobs like these can be profit generators for corporations in our economy. However, quality care requires well-qualified care-givers, and a relatively low ratio of care-givers to those cared for. The families who need this care often have limited resources to pay for it. And care-givers deserve a living wage. The numbers do not add up, to allow all three of these constraints to be satisfied at the same time.[52] For the employer to make a profit, some combination of quality of care, affordability, and living wages must be sacrificed.

---

[50] B. R. Rosso, K. H. Dekas & A. Wrzesniewski. On the meaning of work: A theoretical integration and review. *Research in Organizational Behavior* **30**: 91-127, 2010.

[51] Michael Tomasello. *A Natural History of Human Morality*. Harvard University Press, 2016.

[52] Sally Ho. `Broken' economics for preschool workers, child care sector. *US News*, 8 September 2018. Downloaded 5-20-2019 from https://www.usnews.com/news/business/articles/2018-09-08/broken-economics-for-preschool-workers-child-care-sector.

There are many other jobs that fit this description of being meaningful for the worker, valuable for society, but not supportable as corporate profit centers. Education is a sector with great unmet needs for teachers, aides, managers, counselors, and support staff in preschool, tutoring and mentoring during primary, secondary, and post-secondary schooling, adult and professional education, and other areas. Emergency services, environmental and infrastructure care and development, and medical care and services could all be expanded. Certain tasks, for example care of a small neighborhood park, could conceivably be automated, but having it done by a dedicated community member would result in the job being done at least as well, but would also provide meaningful work for a member of the community. These jobs require subsidies, but as we have seen the wealth of society continues to grow, so the resources for these subsidies exist.

Rather than try to enumerate such jobs, one would hope for a market-based entrepreneurial mechanism that would reward individuals for creating and maintaining such jobs. This mechanism could not be based entirely on profit, but would use a market-based mechanism to effectively allocate society's subsidy for such work.

These three pieces of the puzzle are promising aspects of a way to use the wealth of society for the benefit of the members of society, especially the human members. Making these three pieces fit together will be a challenge, most especially the political task of channeling the resources created by increased automation to the creation of the new jobs the society needs.

8 Conclusions

Ethics is how a society encourages its individual members to interact in positive-sum (cooperative) ways, rather than negative-sum (exploitative) ways, so the interactions strengthen rather than weaken the society as a whole. Ethics accomplishes this goal by encouraging trustworthy behavior by individuals, which earns trust by others, which is necessary for cooperation.

Over centuries, our society has accumulated many different situation-specific ethical principles and social norms that we count on to make our lives together safer and more effective.[53] We individuals use concepts like virtue, duty, utility, etc., to learn, understand, and teach ethical principles. These are the concrete connections from individual ethics, to trustworthiness, to trust, to cooperation, to positive-sum outcomes.

We need to understand what social norms we trust, how trusting them increases positive-sum outcomes for society as a whole, how those norms are represented as knowledge in the minds of individual agents (human and non-human), and how they are applied by agents when making plans and deciding how to act.

This essay has considered examples illuminating three different aspects of ethics from a computational modeling perspective. First, autonomous vehicles are individually embodied intelligent systems that act as members of society. The ethical knowledge needed by such an agent is not how to choose the lesser evil when confronted by a Deadly Dilemma, but how to

[53] Russ Shafer-Landau, editor. *Ethical Theory: An Anthology*. Wiley-Blackwell, second edition, 2013.

recognize the upstream decision point that makes it possible to avoid the Deadly Dilemma entirely.

Second, disembodied distributed intelligent systems like Google and Facebook provide valuable services while collecting, aggregating, and correlating vast amounts of information about individual users. Those individual user models earn money for corporations from advertisers who target users with advertisements, but they can be used much more widely. With inadequate controls, these corporate systems can invade privacy and do substantial damage through either correct or incorrect inferences.

Third, acceptance of the legitimacy of the society by its individual members depends on a general perception of *fairness*: that those who contribute to the success of a collective effort will share in the benefits. Rage about unfairness can be directed at individual free-riders or at systematic inequality across the society.

The promise of a computational approach to ethical knowledge is not simply ethics for computational devices such as robots. Rather, just as artificial intelligence helps us understand cognition, it now also promises to help us understand the pragmatic value of ethics as a feedback mechanism that helps intelligent creatures, human and non-human, live together in thriving societies.

Bibliography

- Joshua Greene. *Moral Tribes: Emotion, Reason, and the Gap between Us and Them*. Penguin Press, 2013.
- Jonathan Haidt. *The Righteous Mind: Why Good People are Divided by Politics and Religion*. Vintage Books, 2012.
- Benjamin Kuipers. How can we trust a robot? *Communications of the ACM*, **61**(3):86–95, 2018.
- Patrick Lin, Keith Abney, & George A. Bekey, editors. *Robot Ethics: The Ethical and Social Implications of Robotics*. MIT Press, 2012.
- Steven Pinker. *Enlightenment Now: The Case for Reason, Science, Humanism, and Progress*. Viking, 2018.
- Peter Singer. *The Expanding Circle: Ethics, Evolution, and Moral Progress*. Princeton University Press, 1981.
- Michael Tomasello. *A Natural History of Human Morality*. Harvard University Press, 2016.
- Wendell Wallach & Colin Allen. *Moral Machines: Teaching Robots Right from Wrong*. Oxford University Press, 2009.
- Robert Wright. *Nonzero: The Logic of Human Destiny*. Pantheon, 2000.
- Shoshana Zuboff. *The Age of Surveillance Capitalism*. Public Affairs, New York, 2019.