CHAPTER 2

THE ETHICS OF THE ETHICS OF AI

THOMAS M. POWERS AND
JEAN-GABRIEL GANASCIA

Introduction

The broad outlines of the ethics of AI are coming into focus as researchers advance the state of the art and more applications enter the private and public sectors. Like earlier technologies such as nuclear fission and recombinant DNA, AI technologies will bring risks and rewards for individuals and societies. For instance, the safety of pedestrians in the path of autonomous vehicles, the privacy of consumers as they are analyzed as data subjects, and the fairness of selection procedures for loan or job applicants—as they are (algorithmically) "scrutinized"—will increasingly be of concern. Those concerns will affect societies as we grapple with the moral and legal status of these new artificial agents, which will increasingly act without direct human supervision. The risks are largely seen as justifying the rewards, and the latter are expected to be significant indeed. Economic forecasts tout robust and relatively certain revenue growth and productivity gains from AI for the next few decades,[1] yet at the same time increased unemployment is expected as industrial labor markets shrink due to rapid AI outsourcing of skilled and unskilled labor. On a more global level, AI will continue to transform science and engineering, but it can also be used to afford leisure and expand knowledge in the humanities.[2] When combined with efficient data-gathering techniques and breakthroughs in genetics, nanoscience, and cognitive science, AI will almost certainly entice

[1] Philippe Aghion, Benjamin F. Jones, and Charles I. Jones, "Artificial Intelligence and Economic Growth," in *The Economics of Artificial Intelligence: An Agenda*, ed. Ajay Agrawal, Joshua Gans, and Avi Goldfarb (Chicago: University of Chicago Press, 2019), 237–82.

[2] Jean-Gabriel Ganascia, "Epistemology of AI Revisited in the Light of the Philosophy of Information," *Knowledge, Technology, and Policy* 23 (2010): 57–73, accessible at: https://doi.org/10.1007/s12130-010-9101-0.

us to effect a greater mastery of our planet. Perhaps AI will first pass through a stage of attempts, via surveillance, policing, and militarization, to also master other human beings.

Faced with this panoply of ethical concerns, which implicate fundamental human rights (privacy, security, equal opportunity), ethical principles (fairness, respect), and equitable distributions of burdens and benefits, it may be useful first to ask: How ought we to approach the ethics of AI? Or, in other words, what are the ethics of the ethics of AI? The preceding account suggests that issues might be engaged on individual, social, and global levels. To be sure, ethicists have begun to make progress on ethical concerns with AI by working within a particular level, and through approaches (deontological, consequentialist, virtue ethics, etc.) common to other fields of applied ethics. Scholarship in machine ethics, robotic ethics, data science ethics, military ethics, and other fields is generating interest from within and without academia. The ethics of AI may be a "work in progress," but it is at least a call that has been answered.

But will this be enough? The thesis of the present chapter is that the common approaches may not be sufficient, primarily due to the transformational nature of AI within science, engineering, and human culture. Heretofore, ethicists have understood key ethical concepts, such as agency, responsibility, intention, autonomy, virtue, right, moral status, preference, and interest, along models drawn almost exclusively from examples of human cognitive ability and reasoned behavior. Ethicists have "applied" ethics accordingly with these conceptual tools at hand. Artificial intelligence will challenge all those concepts, and more, as ethicists begin to digest the problem of continued human coexistence with alternate (and perhaps superior) intelligences. That is to say, AI will challenge the very way in which we have tried to reason about ethics for millennia. If this is correct, novel approaches will be needed to address the ethics of AI in the future. To go further and implement ethics in AI, we will need to overcome some serious barriers to the formalization of ethics.

Further complicating factors in doing the ethics of AI concern epistemic issues, broadly speaking. First, we (ethicists) generally learn of AI applications only after they appear, at which point we attempt to "catch up" and possibly alter or limit the applications. This is essentially a rearguard action. The time lag owes to the fact that ethicists are not in the business of predicting the emergence of technologies. While it would be good if we could figure out the ethics of a technology prior to it being released in the marketplace or public sphere—if we could do "anticipatory ethics"[3]—the necessary predictive skill would not be the domain of ethics. Further, when ethicists *do* try to predict the trajectory of a new technology into future applications in order to critique it, they often get the trajectory wrong. This overestimation of future technological/ethical problems leads some ethicists to become (amateur) futurists, and these futurists often spend an inordinate amount of time worrying about technological applications that will never come to pass.

Second, the epistemic complications of AI turn on the fact that AI itself is changing what we know, especially in the realm of science. Computational data science (CDS),

[3] Philip A. E. Brey, "Anticipatory Ethics for Emerging Technologies," *Nanoethics* 6:1 (2012): 1–13.

which includes "big data" science and other discovery-based techniques, adds immensely to the body of accessible information and correlations about the natural and social worlds, thus changing how scientists think about the process of inquiry. Computational data science calls into question whether this new knowledge really adds to our human scientific understanding. Since many ethical analyses depend on scientifically derived knowledge—especially knowledge of social facts and relations—we are placed in a difficult epistemic position. Whether one conceives of the body of knowledge as a coherentist "raft" or as a foundationalist "pyramid,"[4] the expansion of knowledge due to AI seems to be an epistemic gift, and at the same time we cannot fully understand what we are really getting.

Our goal in the following reflections is not to resolve or even attempt to analyze specific ethical issues that arise with AI. Rather, we will survey what we believe are the most important challenges for progress in the ethics of AI. At the present moment, there are many AI applications that are driving the interest in ethics; among them are autonomous vehicles, battlefield (lethal) robots, recommender systems in commerce and social media, and facial recognition software. In the near future we may have to grapple with disruptions in human social and sexual relationships caused by androids or with jurisprudence administered primarily by intelligent software. The developments in AI—now and in the foreseeable future—are sufficiently worrisome such that progress in the ethics of AI is in itself an ethical issue.

The discussion of these challenges incorporates longstanding philosophical issues as well as issues related to computer science and computer engineering. We leave it to the reader to pursue technical details of both philosophical and scientific issues presented here, and we reference the background literature for such inquiries. The challenges fall into five major categories: conceptual ambiguities, the estimation of risks, implementing machine ethics, epistemic issues of scientific explanation and prediction, and oppositional versus systemic ethics approaches.

Conceptual Ambiguities

Research in ethics and in AI, respectively, involves distinct scholarly communities, so it is not surprising that terminological problems arise. Key concepts in contemporary (philosophical) ethics also appear in the AI literature—especially concepts such as agent, autonomy, and intelligence—though typically ethicists and AI experts attach different meanings to these terms. In this section, we explain standard meanings that attach to these three polysemous concepts in both fields. While we cannot hope to dissolve the ambiguities in favor of one or another meaning, we want to draw attention to them as sources of potential problems within the ethics of AI.

[4] Ernest Sosa, "The Raft and the Pyramid: Coherence versus Foundations in the Theory of Knowledge," *Midwest Studies in Philosophy* 5:1 (1980): 3–26.

Agent

Central to modern AI since the 1980s, the notion of an agent—and one that is supposed to be "intelligent"—has often been seen as the main unifying theme of the discipline. That is particularly apparent in the renowned manual on artificial intelligence by Stuart J. Russell and Peter Norvig, *Artificial Intelligence: A Modern Approach*, which defines AI "as the study of agents that receive percepts from the environment and perform actions."[5] The theme is repeated in the classical "human problem solving" account in Alan Newell and Herbert A. Simon's *Human Problem Solving*,[6] also published in Newell's work in "The Knowledge Level,"[7] and in the widely used notion of multi-agent systems (MAS) that refers to systems composed of a plurality of agents interacting together. In the context of AI, the notion of an agent is closely related to its meaning in economics or in cognitive sciences, since all these terms characterize entities that act. More precisely, following Russell and Norvig, we can say that an AI "agent implements a function that maps percept sequences to actions." Within this definition, the structure of actions is reduced to their mechanical consequences, while their objectives—the goals the agent pursues or, in more philosophical terms, the intentions—are not specified. Those are given from outside, which means that artificial agents do not initiate actions; they are not aware of what they do when acting.

In philosophy, an agent intends (upon reflection) its actions. It is aware of the selection of intentions, and it initiates actions based on them. In other words, artificial agents (for philosophy) do not have agency.

The differences between these two conceptions of agents—the technical one in AI, economics, and psychology as well as the philosophical one—have important consequences from an ethical point of view. Obviously, since an AI agent lacks true proper goals, personal intentions, or real freedom, it cannot be considered to be responsible for its actions, in part because it cannot explain why it behaves in such and such a way and not in other ways. This is not so with the notion of "agent" as understood in its philosophical sense, where an explanation (or an accounting) of action can be expected. This issue has been widely debated in the philosophical community, for instance, in connection with Daniel Dennett's notion of an "intentional system,"[8] which can be used to describe computers to which people ascribe intentions, desires, and beliefs by calling them *intentional agents*.[9] However, even in that case, Dennett clearly specifies that what he calls the "intentional stance" is only a prerequisite for the "moral stance" to which it

[5] Stuart J. Russell and Peter Norvig, *Artificial Intelligence: A Modern Approach*, 3rd ed. (Upper Saddle River, NJ: Prentice-Hall, 2010).

[6] Alan Newell and Herbert A. Simon, *Human Problem Solving* (Englewood Cliffs, NJ: Prentice-Hall, 1972).

[7] Alan Newell, "The Knowledge Level," *Artificial Intelligence* 18 (1982): 87–127.

[8] Daniel C. Dennett, "Intentional Systems," *Journal of Philosophy* 68:4 (1971): 87–106.

[9] Daniel C. Dennett, *The Intentional Stance* (Cambridge, MA: MIT Press, 1987).

cannot be fully assimilated.[10] In other words, a "moral agent" has to be an "intentional system," while there are many "intentional systems," like artificial agents, that are not "moral agents."

Autonomy

The adjective "autonomous" and the concept of autonomy to which it is connected have been widely employed in the last few years to characterize systems that behave without human intervention. More precisely, a device is said to be autonomous if there exists a sequence of cause-effect relations—from the capture of information by sensors to the execution of an action—without the intervention of any human being. Referring to this definition, AI researchers currently speak of autonomous cars, weapons, and (perhaps in a more frightening way) of "lethal autonomous weapon systems" (also referred to as LAWS). In these usages, it is very difficult to distinguish autonomy from automaticity, since in both cases the relevant behavior corresponds to entities that act by themselves, which clearly corresponds to the etymology of *automaton*: αυτο (self) + ματος (movement). However, not only does the etymology of autonomy—αυτο (self) + νομος (law)—differ from that of automaticity, but its usual meaning, at least for philosophers, designates an entity able to define by itself its own laws or rules of behavior, while in the case of an automaton these rules are given or imposed from outside. Originally, the adjective "autonomous" described a political entity (e.g., a sovereign city, kingdom, or state), which decided by itself its constitution and its laws. This meaning survives in the granting of limited self-rule in the several "autonomous regions" of various nation-states. Following the philosophers of the Enlightenment, in particular Rousseau and Kant, this meaning of autonomy has been extended to human beings. Here it denotes an ideal situation in which individuals would decide their maxims of conduct for themselves without being commanded by kings, presidents, or others. So, in a way, an autonomous being that obeys its own rules will choose them by itself, and thus will reflect on what it will do, while an automaton acts by obeying rules imposed on it and without reflection.

To see why the semantics matters, let us consider an example. Suppose we want an autonomous vehicle to drive us safely to the destination that we have indicated. For instance, if we want to go to the swimming pool, and we clearly indicate to the car that this what we want, we expect such a technology to adopt that specified goal. Now, let us assume that the car is autonomous (according to the philosophical understanding), i.e., that it decides by itself, and not following a person's order, what will be its goal and rule of conduct. It may choose to make an appointment for you at the dentist (perhaps in a paternalistic way), or drive you to the movie theater because the parking there looks to be more comfortable for it. As a consequence, a "real" autonomous car is above all

[10] Daniel C. Dennett, "Mechanism and Responsibility," in Ted Honderich (ed.), *Essays on Freedom of Action* (London: Routledge & Kegan Paul, 1972), 157–84.

somewhat unpredictable for the person who is being conveyed by it, and consequently it is not so desirable as a mode of transportation!

Worse still, imagine a "real" (philosophically) autonomous weapon that would choose by itself who it would target. This would be a nightmare not only for civilians and noncombatants but also for military personnel who need, first and foremost, weapon systems that they can fully control and trust. From this point of view, it is quite unlikely that a military would develop "real" autonomous weapons, even though autonomous weapons that fit the AI or engineering definition seem quite desirable.

In many philosophical traditions, agency and autonomy are properties of adult, rational beings or moral persons who have the ability to choose and regulate their own behaviors. Agency and autonomy are necessary conditions of responsibility. In AI an agent is a piece of software within a larger computer system that performs a function on behalf of a user or another software agent. An autonomous agent in AI is a piece of software that functions more or less continuously without the direct intervention of a user. In AI, the concepts of agent and autonomy are used without any obvious connection to responsibility. As a result of these conceptual differences, it is important to recognize that a (philosophical) autonomous agent acts on its own behalf, and has the ability to "intervene" in its own behavior (at the least), while a (software) AI autonomous agent does not itself have a concept of "its own behalf." This is not to say that it is inconceivable that someday there will be software agents that act absolutely without human intervention and on their own behalf. Perhaps then it will make sense to attribute responsibility to them for their actions. But the point is that, with the AI agents we now have, this is not the case. Nonetheless, there are still ethical issues that arise when AI agents act on the behalf of other users or software agents, and also when they act (relatively) independently of human intervention.

Intelligence

Though philosophical studies of intelligence, going back to Vico's work in the eighteenth century, considered it to be a distinctively human ability, it is now acknowledged that intelligence can have other instantiations. Because it plays such an important role both in AI and in the public imagination of computation in general, the concept of intelligence needs to be clarified. In early modern philosophy, intelligence was typically interchangeable with understanding and indicated an ability to comprehend or grasp aspects of an internal or external reality. In contemporary philosophical usage, intelligence has largely been supplanted by the concept of mind. In the natural and social sciences, especially in psychology, intelligence denotes cognitive abilities that are susceptible to measurement—for instance, via an intelligence quotient that aggregates the results of different tests in order to grade the relative abilities of people in a population.

The technical meaning of "intelligence" in AI—one that assumes that we can engineer intelligence—derives from its significance in psychology. The proposal of the Dartmouth Summer Research Project on Artificial Intelligence (written mainly by John

McCarthy and Marvin Minsky) contains in its introduction the central motivating claim of AI: "The study [of Artificial Intelligence] is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it."[11] Intelligence is here conceived as a set of mathematically describable cognitive functions, which AI aims to model and then simulate with machines.

Despite this narrowing of intelligence into a technical concept, it has taken on a meaning in both the public imagination and the marketing literature of some IT companies, along with a significance that includes a mixture of very different capacities: will, consciousness, reflection, and even an aptness to perceive and feel emotions. Unfortunately, discussions about the intelligence of AI systems are often an admixture of popular, philosophical, and scientific conceptions.

Closely connected to intelligence in work on the philosophy of mind is the (philosophical) notion of consciousness. One standard assumption in philosophy is that all intelligent entities have consciousness as the "backdrop" or "framework" in which intelligence happens, as it were. Though some philosophers such as David Chalmers see in consciousness a "hard problem,"[12] which suggests that it may never be integrated into the physical sciences, consciousness is sometimes employed by writers in AI to characterize a possible capacity of future intelligent systems. But unlike in philosophy, there is no assumption of an intelligent computer's "first-person perspective" nor a "having" of computational states that are equivalent to mental states that philosophers call "qualia," that is, "what it is like" to have a particular awareness (e.g., seeing the red apple). A middle-ground notion of consciousness has been suggested, according to which a machine would behave as though it were conscious if it had (1) global availability of relevant information (access to an "internal global workspace") and (2) self-monitoring ("reflexive representation").[13] Here we see the return of Dennett's intentional stance, with a measure of behaviorism thrown in.

To conclude this section on the conceptual ambiguities that arise in ethical debates around AI, let us consider two broadly used terms in the field: "intelligent agent" and "autonomous agent." Taking into account what we have said about the philosophical meanings of these terms, they seem to resemble the famous Lichtenberg knife (which lacks a blade and a handle), since the "autonomous agents" are neither autonomous nor agents (for the philosophers), and likewise "intelligent agents" are neither intelligent nor agents.

[11] John McCarthy, Marvin Minsky, Nathaniel Rochester, and Claude E. Shannon, "A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence" (1955), accessible at: http://raysolomonoff.com/dartmouth/boxa/dart564props.pdf.

[12] David J. Chalmers, "Facing up to the Problem of Consciousness," *Journal of Consciousness Studies* 2 (1995): 200–219.

[13] Stanislas Dehaene, Hakwan Lau, and Sid Kouider, "What Is Consciousness, and Could Machines Have It?" *Science* 358:6362 (2017): 486–92.

Risk: Overestimation and Underestimation

Partly due to the aforementioned ambiguities, and partly to current social demand driven by popular media,[14,15,16,17] which overemphasize the "dangers" of AI, estimations of the risks of AI suffer from both excess and deficiency. On the side of excess, the presumed dangers include allegedly autonomous AIs that operate without any human control, the weaponization of AI globally, and the development of an AI that would "choose its own ends." The popular media as well as some AI experts have fallen into the confusion over agency and autonomy in machines, as indicated earlier, and may become fixated on speculative risks. One example is the recent focus on driverless cars and the claim that they will introduce potentially unsolvable "trolley problems" into the application of these AI technologies. On the side of deficiency, there are AI systems that present real (but underestimated) risks now. For instance, using AI techniques, deepfake software synthesizes fake human pornographic videos that combine and superimpose an existing person's face on a prerecorded video with a different body, so that this person seems to do or say things that he/she never did. Another overlooked application of AI comes in facial recognition and recommending techniques that have been implemented in China to give a "reputation score." The system automatically identifies minor law infractions by citizens, for instance crossing the road at the green light, and aggregates them. Such examples suggest that identity, sexual orientation, consumer tendencies, and the like will all be subject to AI tools. In this section, we discuss the ethical implications of under- and overestimation of AI risks.

Overestimations and Existential Threats from AI

Among the current overestimations of AI, some critiques revisit earlier fears about technology in general. By mimicking human behaviors and abilities, AI, it is feared, creates (or may soon create) artificial human beings and, in so doing, will attempt to "play" or

[14] Joel Achenbach, "Driverless Cars Are Colliding with the Creepy Trolley Problem," *Washington Post* (December 29, 2015), accessible at: https://www.washingtonpost.com/news/innovations/wp/2015/12/29/will-self-driving-cars-ever-solve-the-famous-and-creepy-trolley-problem/.

[15] Joel Achenbach, "The A.I. Anxiety," *Washington Post* (December 27, 2015), accessible at: http://www.washingtonpost.com/sf/national/2015/12/27/aianxiety/.

[16] Patrick Lin, "The Ethics of Autonomous Cars," *The Atlantic* (October 8, 2013), accessible at: http://www.theatlantic.com/technology/archive/2013/10/the-ethics-of-autonomous-cars/280360/.

[17] Henry A. Kissinger, "How the Enlightenment Ends: Philosophically, Intellectually—in Every Way—Human Society Is Unprepared for the Rise of Artificial Intelligence," *The Atlantic* (June 2018), accessible at: https://www.theatlantic.com/magazine/archive/2018/06/henry-kissinger-ai-could-mean-the-end-of-human-history/559124/.

"challenge" God as the Supreme Maker. If this were the case, AI would commit, at the least, a symbolic transgression. As an illustration, consider both the enthusiasm and fear that attended the public unveiling of Japanese roboticist Hirochi Ishiguro's Geminoids.[18] By so closely approximating his own appearance with a robot, Ishiguro invited a comparison to the myth of Pygmalion, who falls in love with his statue Galatea. Nonetheless, Ishiguro's robot was not at all autonomous; it was remotely controlled. In the same way, the robot "Sophia," developed by the company Hanson Robotics, received "citizenship" in Saudi Arabia after her speech at a United Nations meeting. The speech was not automatically generated by "Sophia" herself but prerecorded by an organic human female.

Instances of "overselling" of scientific results seem also to be subject to amplification when AI techniques are involved. Psychologists recently published claims that a deep neural network has been trained to better detect sexual orientation from facial images than can humans.[19] The ethical issues here are multiple. It is unclear that AI is in fact capable of such results, given the assumption that sexual orientation is fixed by genetics. That uncertainty notwithstanding, the use of such techniques could be damaging for homosexuals, regardless of the robustness of results. Likewise, there is considerable interest in brain-computer interfaces (BCI), which are supposed to directly plug a brain (or should we say, a mind?) into a computer network without pain or effort. These alleged "mind reads" have drawn the attention of famous technologists such as Mark Zuckerberg.[20] However, the current state of the art does not warrant belief in a generic human-machine interface, though research has shown that stroke patients may regain motor control of a limb through such interfaces.[21] These doubts notwithstanding, Neuralink, a firm founded by Elon Musk, offers another illustration of the allure of a direct connection between our mortal minds and the (immortal) digital world. This company aims at developing plug-in chips in our skull to increase our cognitive abilities and, more specifically, our memory in order to "save the human race" against AI. These hopes are a double overestimation of AI: the first is that AI will constitute an existential threat for humanity; and the second is that AI technology can be used to avoid such a disaster. According to Musk, one difficult task when merging our mind to the digital is that "it's mostly about the bandwidth, the speed of the connection between your brain and the digital version of yourself, particularly output."[22] However, contemporary

[18] Erico Guizzo, "The Man Who Made a Copy of Himself," *IEEE Spectrum* 47:4 (April 2010): 44–56.

[19] Yilun Wang and Michal Kosinski, "Deep Neural Networks Are More Accurate Than Humans at Detecting Sexual Orientation from Facial Images," *Journal of Personality and Social Psychology* 114: 2 (2018): 246–57, accessible at: http://dx.doi.org/10.1037/pspa0000098.

[20] Noam Cohen. "Zuckerberg Wants Facebook to Build a Mind-Reading Machine," *WIRED* (April 2019), accessible at: https://www.wired.com/story/zuckerberg-wants-facebook-to-build-mind-reading-machine/.

[21] Society for Neuroscience, "Potential Brain-Machine Interface for Hand Paralysis: Combining Brain Stimulation with a Robotic Device Could Help Restore Hand Function in Stroke Patients,"*Science Daily* (January 15, 2018), accessible at: www.sciencedaily.com/releases/2018/01/180115151611.htm.

[22] Nick Statt, "Elon Musk Launches Neuralink, a Venture to Merge the Human Brain with AI," *The Verge* (March 27, 2017), accessible at: https://www.theverge.com/2017/3/27/15077864/elon-musk-neuralink-brain-computer-interface-ai-cyborgs.

neurosciences have no idea of the cortex's internal code, which means that the issue of the "link" is not so straightforward. Further, if such devices were really in service and plugged into our brains, the owners of these technologies could always load whatever information they wanted into a "linked" mind, which would give them considerable power over us.

Besides these specific examples of AI technology hopes and fears, there exist other overestimations of AI progress that might be called "existential" in that they purportedly threaten the future of humanity. Among them, some are of particular importance because they claim that humankind will very soon become obsolete. In 1956 Günther Anders announced this thesis in a book that would eventually be translated as *The Obsolescence of Man*.[23] This pessimistic view would be repeated by the famous astrophysicist Stephen Hawking and the theoretical physicist and Nobel laureate Frank Wilczek. A slightly less pessimistic view is that humans will join with machines in a kind of hybrid, which would then offer, at the least, an extension of life or possibly immortality. Proponents of this last view are scientists such as Ray Kurzweil, philosopher Nick Boström, and Musk.

The obsolescence and replacement views are sometimes based on the Singularity hypothesis and the possibility of superintelligence. One of the first expressions of these ideas goes back to 1962 when it was proposed by British statistician Irvin John Good,[24] who had worked with Alan Turing during World War II. Good discussed the possibility of an "intelligence explosion" that would follow the development of "ultra-intelligent machines," themselves able to build more intelligent machinery. The Polish mathematician Stanislaw Ulam and science fiction writers, including Isaac Asimov, are also credited with inventing the idea in the 1950s that a "Singularity" could be the consequence of the considerably accelerating progress of computer technology.[25]

Science fiction novelist Vernor Vinge popularized the idea in an essay entitled "The Coming Technological Singularity."[26] He argued that within less than thirty years, the progress of information technology would allow the making of a superhuman intelligent entity that would dramatically change the status of humankind. In particular, the connection of humans to machines and their mutual hybridization would allow us to considerably increase our intelligence, our lifespan, and capacities of all kinds. The key idea is that the acceleration of technological progress would suddenly and irreversibly alter the regime of knowledge production, creating technological developments beyond any hope of control.

[23] Günther Anders, *Die Antiquiertheit des Menschen Bd. I: Über die Seele im Zeitalter der zweiten industriellen Revolution*. (Munich: C. H. Beck, 2018).

[24] Irving J. Good, "Speculations Concerning the First Ultraintelligent Machine," *Advances in Computers* 6 (1966): 31–88.

[25] Isaac Asimov, "The Last Question," *Science Fiction Quarterly* 4:5 (Nov. 1956).

[26] Vernor Vinge, "The Coming Technological Singularity: How to Survive in the Post-Human Era," in *Vision 21: Interdisciplinary Science and Engineering in the Era of Cyberspace* (Cleveland: NASA Lewis Research Center, 1993), 11–22.

More recently, technologists like Ray Kurzweil,[27] Hans Moravec,[28] Hugo de Garis,[29] Kevin Warwick,[30] Bill Joy,[31] and even philosophers such as Nick Boström and Julian Savulescu[32] have theorized a future where the technological Singularity was supposed to play a major role. There are differences among all of these writers; some consider new plagues generated by the development of computing power while others proclaim the end of humankind and the emergence of a new species. What is common to their views is the rather credulous leap to the conclusion that the Singularity is a coherent scientific eventuality.

Despite its popularity, the main idea of the Singularity is quite dubious. In fact, it appears just to be an inference from the exponential increase of computing power characterized by Moore's law, which will somehow lead to ultraintelligent machines. However, Moore's law—put forward in 1965—is an empirical description of the evolution of hardware. It describes the increase in computing speed, along with an exponential diminution of the cost of storage devices, as borne out by historical evidence. It has held, more or less, for sixty years now. Moore's law makes an inductive prediction; it is not based on the rigorous foundations of computer science. Its main scope was originally economical, not scientific. As a consequence, there are good reasons to doubt that it will hold indefinitely. In addition, the "amount" of intelligence—a strange notion assumed by advocates of the Singularity—can neither be measured by the frequency of a computer's processing speed nor by the quantity of bits that can be stored in electronic devices. Since its beginning, AI progress has been related to algorithms, to statistics, to mathematical probability theory, and to knowledge representation formalisms or to logic, but not to computing power. And though the efficiency of modern computers renders possible the implementation of parallel algorithms on huge quantities of data, there is no assurance that these developments get us any closer to the Singularity.

Underestimation of AI Risks

Along with these abundant overestimations of AI capacities, which are supposed to be either excessively beneficial for humankind or excessively maleficent, many predatory applications of AI techniques are partly ignored, or at least their potential harm is

[27] Ray Kurzweil, *The Singularity Is Near: When Humans Transcend Biology* (New York: Penguin Books, 2006).

[28] Hans Moravec, "When Will Computer Hardware Match the Human Brain?" *Journal of Evolution and Technology* 1 (1998).

[29] Hugo de Garis, *The Artilect War: Cosmists vs. Terrans: A Bitter Controversy Concerning Whether Humanity Should Build Godlike Massively Intelligent Machines* (Palm Springs, CA: ETC Publications, 2005).

[30] Kevin Warwick, *March of the Machines: The Breakthrough in Artificial Intelligence* (Champaign: University of Illinois Press, 2004).

[31] Bill Joy, "Why the Future Doesn't Need Us," *WIRED* 8 (2001): 1–11.

[32] Nick Bostrom and Julian Savulescu, eds., *Human Enhancement* (Oxford: Oxford University Press, 2008).

scarcely noticed. What we here characterize as "underestimations" of AI risks are just as problematic from an ethical point of view as are overstatements of nonexistent threats. Here we consider a few of these neglected "underestimations" of some AI techniques.

Many famous people seem to fear LAWS—lethal autonomous weapon systems—and propose an official multilateral ban to stop research and military applications in this area.[33] Nonetheless, there are serious doubts whether fully autonomous weapons will ever be developed, since, as mentioned above, what armies need are robust and trustworthy weapons.[34] However, as revealed by "The Drones Papers,"[35] information technologies incorporating many AI components have been used in the drone war in Afghanistan to target supposed terrorists. Drones and more generally unmanned weapons are not autonomous, since they are remotely controlled, but the choice of objectives is done partially automatically, based on informational indices. For instance, conversations or phone localizations have provided targets, and these military uses of AI can contribute to considerable collateral damage (and probably already have).

A second example concerns the state use of facial recognition techniques. Without proper safeguards, these techniques can infringe on individual rights as well as threaten the "dignity of the person" by constant surveillance and guilt by association. They could be used to track and record movement of individuals, especially in urban environments with high density of population. It has been reported that China is now using these techniques to track the minority Uighur population,[36] and facial recognition in China could be combined with their more far-reaching "social credit system" for the entire country.[37] For security reasons, some cities in other countries, for instance the city of Nice in France, plan to use facial recognition to detect suspects of terrorism. We should worry that once in place, the scope of application of such AIs would be extended to all citizens.

A further underestimated risk involves machine learning to predict risk for insurers and to apportion the risk by individualizing insurance premiums. Here there are at least two perverse effects. The first concerns the opacity of the decision criteria, which are not given to clients because most of the time they are not explicit, due to the deep learning techniques on which they are based. Some researchers have become aware of problems with opacity and have tried to introduce explainable AI systems. Explanation is crucial in order to earn public confidence, since without explanation the decisions of the insurance company could be totally arbitrary and based on marketing factors more than

[33] Future of Life Institute, "Autonomous Weapons: An Open Letter from AI & Robotics Researchers," published online (July 28, 2015), accessible at: https://futureoflife.org/open-letter-autonomous-weapons/.

[34] Jean-Gabriel Ganascia, Catherine Tessier, and Thomas M. Powers, "On the Autonomy and Threat of 'Killer Robots,'" *APA Newsletter on Philosophy and Computers* 17:2 (2018): 3–9.

[35] The Intercept, "The Drone Papers," published online (October 15, 2015), accessible at: https://theintercept.com/drone-papers/.

[36] Paul Mozur, "One Month, 500,000 Face Scans: How China Is Using A.I. to Profile a Minority," *New York Times* (April 14, 2019), accessible at: https://www.nytimes.com/2019/04/14/technology/china-surveillance-artificial-intelligence-racial-profiling.html.

[37] Rachel Botsman, "Big Data Meets Big Brother as China Moves to Rate Its Citizens," *WIRED* (October 21, 2017), accessible at: https://www.wired.co.uk/article/chinese-government-social-credit-score-privacy-invasion.

on risk.[38] But the second perverse effect would be to change the original nature of insurance, which relies on mutualizing (pooling) risks, and consequently to weaken solidarity and a sense of community.

A final underestimated risk of AI to be considered here concerns predictive justice, which aims at establishing sanctions according to the risk of repeat offenses of the law. Depending on the criteria that are used, these applications could not only be unjust but also deny the relevance of redemption and contrition. In addition, this raises fundamental questions about the nature of juridical sanction, which in principle has to be based on actual infringement of laws and not on potential offense. As in the short story "The Minority Report" (1956) by Philip K. Dick and the film adaptation *Minority Report* directed by Steven Spielberg (2002), this AI application could lead to the punishment of persons guilty of a precrime, that is to say, of a crime that has not yet been committed but that in all probability will be.

Implementing Ethics

Making Machines Moral

Undoubtedly, it would be tempting to introduce human values in machines to make them moral, which means to make them behave in accordance with criteria of moral behavior generally, or, for the deontologist, to act only according to duty. We might then ponder the distinction, attributed to Kant, between acting merely in conformity to duty versus acting from a sense of it, which the good will alone achieves. However, since a machine does not determine its own ends or goals of action, but acts on goals given to it from outside, invoking will—that is, diving errantly into machine motivations—would seem foolish. Thus, we shall only consider here the ability of a machine to *behave* morally, without invoking its moral motivations.

In the past few years, some AI researchers[39,40,41,42,43,44] have attempted to theorize intelligent agents that appeal to ethical considerations when choosing the actions they

[38] Cathy O'Neil, *Weapons of Math Destruction* (New York: Crown Publishers, 2016).

[39] Fiona Berreby, Gauvain Bourgne, and Jean-Gabriel Ganascia, "Event-Based and Scenario-Based Causality for Computational Ethics," in *Proceedings of the 17th Conference on Autonomous Agents and MultiAgent Systems*, (Richland, South Carolina: International Foundation for Autonomous Agents and Multiagent Systems, (2018): 147–55.

[40] Selmer Bringsjord, Konstantine Arkoudas, and Paul Bello, "Toward a General Logicist Methodology for Engineering Ethically Correct Robots," *IEEE Intelligent Systems* 21:4 (2006): 38–44.

[41] Jean-Gabriel Ganascia, "Modelling Ethical Rules of Lying with Answer Set Programming," *Ethics and Information Technology* 9:1 (2007): 39–47.

[42] Wendell Wallach, Colin Allen, and Iva Smit, "Machine Morality: Bottom-up and Top-down Approaches for Modelling Human Moral Faculties" *AI & Society* 22:4 (2008): 565–82.

[43] Thomas M. Powers, "Prospects for a Kantian Machine," *IEEE Intelligent Systems* 21:4 (2006): 46–51.

[44] Amitai Etzioni and Oren Etzioni, "Incorporating Ethics into Artificial Intelligence," *Journal of Ethics* 21:4 (2017): 403–18.

perform. This work can be seen as a response to potentially unpredictable behaviors in machines, as when machine-learning techniques build opaque programs from huge quantities of training examples that no human would be able to assimilate. In such situations, not only are machines unable to explain their behavior in terms understandable by humans but also their decisions could produce significant harms. It therefore seems crucial to control machine behaviors to ensure that they conform to shared social norms and values. This section will give an overview of some ways to introduce ethical controls and also will describe their intrinsic limitations. We note that these approaches are quite remote from actual ethical issues related to current applications of AI, but may become more relevant as AI advances.

Modeling Ethical Reasoning

At first sight, it may seem plausible to model ethical systems with AI techniques, since the prescriptions on which such systems are based have been introduced by humans. However, the attempts to model ethical reasoning have shown the huge difficulties researchers face in doing so. The first difficulty comes from modeling deontic reasoning, that is, reasoning about obligations and permissions. The second is due to the conflicts of norms that occur constantly in ethical reasoning. The third is related to the entanglement of reasoning and acting, which requires that we study the morality of the act, per se, but also the values of all its consequences.

To solve the first of these difficulties, concerning the particular nature of rules of duty, some researchers have used deontic logics[45,46] and formalisms inspired by deontic considerations. The second difficulty is approached by the use of techniques that overcome logical contradictions with AI logic–based formalisms,[47] mainly nonmonotonic formalisms (e.g., default logics[48] and answer set programming),[49] which capture aspects of commonsense reasoning. Lastly, the third approach intertwines the logic-based models of ethical reasoning to formalisms called action languages[50] or causal models,[51] which have been designed to give a clear semantics that provide a strong mathematical grounding

[45] Emiliano Lorini, "On the Logical Foundations of Moral Agency," in *International Conference on Deontic Logic in Computer Science*, ed. T. Ågotnes, J. Broersen, D. Elgesem, *Deontic Logic in Computer Science: DEON 2012*, Lecture Notes in Computer Science 7393 (Berlin: Springer, 2012), 108–22.

[46] John F Horty, *Agency and Deontic Logic* (Oxford: Oxford University Press, 2001).

[47] Jean-Gabriel Ganascia, "Non-monotonic Resolution of Conflicts for Ethical Reasoning," in *A Construction Manual for Robots' Ethical Systems*, ed. Robert Trappl (Cham, Switzerland: Springer International Publishing, 2015), 101–18.

[48] Raymond Reiter, "A Logic for Default Reasoning," *Artificial intelligence* 13:1–2 (1980): 81–132.

[49] Michael Gelfond, "Answer Sets" *Foundations of Artificial Intelligence* 3 (2008): 285–316.

[50] Erik T. Mueller, *Commonsense Reasoning: An Event Calculus Based Approach* (Burlington, MA: Morgan Kaufmann, 2014).

[51] Joseph Y. Halpern and Max Kleiman-Weiner, "Towards Formal Definitions of Blameworthiness, Intention, and Moral Responsibility," *Proceedings of the 32nd AAAI Conference on Artificial Intelligence* (2018): 1853–60.

for understanding the consequences of actions. The technical challenge nowadays is to merge these three approaches, that is to say, to create one that is nonmonotonic, that can handle conflicts of norms, and that uses causal models to evaluate the consequences of actions. While there is a general interest in creating such a moral machine (i.e., one that behaves in conformity with the rules of a morality), all these approaches embrace different normative frameworks—such as utilitarianism, egoism (game theory), deontology, and virtue ethics approaches—that must be simulated. The details of the simulations are usually found to be lacking, especially by philosophers. In addition, there are questions about the practical utility of such moral machines as well as the difficulties in implementing them.

Learning Values

Whatever normative framework is used to simulate moral reasoning, the presumption is that it will be based on values that need to be acquired by the machine and that depend on societies and their ethical traditions. Considering the relativity of norms and values on which moral decisions are made, a few attempts[52, 53] have been made to use machine-learning techniques to automatically learn moral values and rules on which machine morality would be based. The popularity and the efficiency of machine learning drives such projects from a technical point of view, even if they can be criticized from an ethical point of view. Since ethics is not just a question of social acceptancy but also of prescriptions that are not based on observations of how people act (i.e., based on conceptions of how they ought to act), the ethics of AI will have to grapple with this basic difference in approaches to ethics.

To make this concern more concrete, consider the highly publicized "Moral Machine Experiment" that gathered attitudes about how autonomous vehicles ought to solve moral dilemmas in various crash-trajectory scenarios where people (variously described) or animals were put at risk, and others were spared.[54] The researchers employed an online experimental platform to crowdsource attitudes by collecting 40 million preferences from millions of persons across 233 different countries. The researchers compared the attitudes of respondents across regions, countries, cultures, religions, and genders. The results suggested that variations in ethical attitudes correlate with deep cultural traits, and perhaps even with adherence to different moral principles.

[52] David Abel, James MacGlashan, and Michael L. Littman, "Reinforcement Learning as a Framework for Ethical Decision Making," in *The Workshops at the 30th AAAI Conference on Artificial Intelligence*, Technical Report WS-16–02 (Palo Alto, CA: Association for the Advancement of Artificial Intelligence, 2016): 54–61.

[53] Max Kleiman-Weiner, Rebecca Saxe, and J. B. Tenenbaum, "Learning a Common-Sense Moral Theory," *Cognition* 167 (2017): 107–23.

[54] Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan, "The Moral Machine Experiment," *Nature* 563 (2018): 59–64.

This is undoubtedly an important result from a social psychology and an empirical-ethics point of view, as it provides evidence of relevant variations in ethical attitudes.

Nevertheless, the researchers seem also to have a normative goal in mind: to introduce these results into the design of autonomous vehicles so that they adapt to local cultures and expectations of the (presumably homogenous) populations where the vehicles will operate. So quite directly the experiment implicates the longstanding issue in ethics about conventionalism and ethical relativity versus the validity of generalizable ethical principles or duties that ethicists might prefer. The authors confront this issue and note that solutions to moral dilemmas provided by ethicists could very well be rejected by the public, and thus might be (in their words) "useless." The lesson here for the ethics of ethics of AI is that there are bound to be approaches to AI ethics that advocate conformity with varying public attitudes. But would ethicists be approving of adultery, for instance, simply because it is widely practiced? When it comes to doing the ethics of AI, should ethicists resist "following the data" and insist on generalizable solutions to moral dilemmas that might strike some publics as "out of touch"? To choose the former "empirical" approach would be to swear off the latter traditional philosophical conception of normativity, but also would allow AI applications to take advantage of machine learning over large datasets. And it is important to note the enthusiasm for machine learning over "big data," which may well influence the development of some ethics of AI.

Intrinsic Limitations

In addition to the controversy over the source of values on which ethical deliberations in AI will be based, another crucial question concerns what constitutes the intelligence of AI agents. As an illustration, consider that the fatal accident of Uber's self-driving car in 2018 in Arizona was not due to faulty sensors but to the decision of Uber, for the sake of the passengers' comfort, to moderate reactions to unidentified obstacles such as leaves or plastic bags. This means that the accident in question was not due to an unethical deliberation but to a fateful judgment about safety versus comfort that had been programmed by engineers.

In a totally different context—that of lethal battlefield robots—Ron Arkin's ethical governor[55] for robot soldiers provides another illustration of hard problems that automatic AI systems will have to face. Arkin proposes to use AI techniques to implement just war theory, the International Laws of War, and a particular operation's Rules of Engagement in a control module called the *ethical governor*. This is supposed to control a robot soldier's decision procedures to make it more ethical than human soldiers, who, under the emotional pressures of battle, often feel anger, fatigue, and desperation and thus behave inappropriately. Among the *jus in bello* rules that need to be implemented

[55] Ronald C. Arkin, Patrick Ulam, and Brittany Duncan, "An Ethical Governor for Constraining Lethal Action in an Autonomous System," Technical Report GIT-GVU-09-02, Georgia Institute of Technology Mobile Robot Lab (2009).

in such situations are the discrimination between military personnel and civilians and the protection of civilians. However, especially in asymmetric conflicts where soldiers do not wear uniforms, such discrimination is very difficult, even for humans. How can we ensure that a robot will correctly discriminate? This is a question of judgment—understood not as juridical or normative judgment but rather as an operation of categorizing objects in a situation from flows of information. Further, the discrimination rule has two exceptions: (1) when human soldiers are disarmed, they can be taken prisoner but must be protected according to international laws; and (2) when civilians take part in hostilities, they become combatants and can be attacked. In both cases, the intelligence of the judgment or categorization precedes the ethical deliberation; in fact, it seems to exhaust it. It appears that the practical problems are not due to difficult ethical deliberations, of which the autonomous vehicle "crash" dilemma is certainly the most popular illustration, but to questions of judgment, which are difficult even for humans.

Epistemic Issues with Ethical Implications: Predictive Science

In recent decades the role of epistemology in ethics has emerged from some traditional concerns of moral or meta-ethical epistemology, that is, issues about the nature of moral knowledge, what counts as evidence for moral claims, and the like. The more recent concerns highlight the simple, practical point that *what* one knows or believes tends to structure one's ethical obligations. Ethical disputes can indeed revolve around the grounds for obligation, but even assuming agreement on the grounds, disputes can also arise concerning the facts that would activate an obligation. For instance, suppose two agents believe in general that saving the planet from environmental ruin is an obligation, but one of them denies that climate change is real and has been deprived of knowledge of it. Then that latter agent is not (practically speaking) obligated to act to save the planet; the agent lacks the motivation because she lacks knowledge. Knowing precedes recognition of an obligation to act.

Artificial intelligence enters the concern about epistemology in ethics in virtue of the fact that AI is an increasingly large "supplier" of scientific information and results—especially in those disciplines identified as practicing Big Data science—and as AI continues to grow in importance for science, our epistemic dependence on AI will only increase. This will be true of descriptions of the natural world, but also of predictions, since they come from data-intensive mathematical models. So another important challenge for the ethics of ethics of AI is how AI is increasingly used to establish scientific facts, and whether those facts can be readily explained either to the lay public or in some cases even to expert scientists themselves. Here we focus on ways in which AI might create a future body of scientific results that will fall short of adding to our scientific understanding. The problem is a peculiar feature of AI in that there can be considerable

generated knowledge (in terms of correlations of data and phenomena), but no commensurate increase in genuine human scientific understanding.

We will use the term computational data science (CDS) to refer to the collection of computationally based scientific techniques, primarily involving AI, that were developed in the late twentieth century to probe our natural and social worlds. These forms of AI rely on other information technologies that generate and store large amounts of data, so CDS proper should be understood as a result of both AI and modern (nonintelligent) data producing and gathering technologies. As American computer scientist Peter J. Denning has written, CDS brought a "quiet but profound revolution" that has transformed science by making new discoveries possible.[56] What is striking about CDS is the presumed agency of "making new discoveries possible," for there is a very clear sense in which *computers* and not humans are now making these scientific discoveries. There is a further concern that the progress of CDS is leaving human scientists behind—almost as though we are becoming adjuncts to the scientific discovery process. This is a serious worry, and here we will characterize some of its aspects concerning (1) the tension between statistical and causal accounts of "associationist" CDS; (2) the notion that scientific understanding (as a broad cognitive phenomenon) is threatened by CDS; and (3) that CDS poses problems for ethics—here considered in two ways: (a) the possibility of new statistical ethical knowledge about individuals, and (b) the application of statistical methods through CDS to decide social policies and interventions in areas such as public health and criminal justice.

These three topics—causal knowledge, scientific understanding, and the use of statistics in ethics—are far from the only philosophical topics that CDS implicates. There are a myriad of ways in which CDS has changed science, and will increasingly change technology as control architectures of robots and AI systems become integrated with real-time "Big Data" results. Likewise, as philosophers of science turn their attention to the philosophy of CDS, there may be many other important investigations to undertake, including the application of CDS to the explanation of consciousness, free will, the status of scientific laws, and so on. An analogy to the present historical moment of CDS is provided by the now-common television "extreme weather" journalism, where a reporter outfitted in rain gear stands on a beach that is in the path of a hurricane, in breathless excitement as the first rains start to fall. We have a good idea of what's coming, it is quite certain to be a deluge, but it would be foolish to think we know in detail what the storm will be.

It is difficult to say when exactly CDS as a revolution begins. Denning cites the work of the Nobel physicist Kenneth Wilson in the 1980s, who developed computational models for phase changes and the direction of magnetic force in materials. Wilson was also a passionate advocate for CDS and lobbied American science-funding agencies to secure more support for the field. These efforts resulted in the High-Performance Communication and Computing (HPCC) Act of 1991 in the United States—in large part

[56] Peter J. Denning, "Computational Thinking in Science," *American Scientist* 105:1 (January–February 2017): 13–17.

through the efforts of former vice-president Al Gore. The HPCC was one reason that Gore infamously claimed that he "invented the Internet"—and thus we might go back further to give credit to the creation of ARPANET as the beginning of CDS. Whenever our starting point, it is clear that CDS includes advances in the science of simulation, which revolutionized fields from aeronautics to theoretical physics to computer modeling for everything from climate change to recidivism rates for human criminal activity, as well as advances in modern biology, bioinformatics, DNA sequencing, systems and synthetic biology, and now even single-nucleotide gene editing. It is safe to say that for any science for which there are large amounts of data that are available, and where computation over those datasets is impractical for human practitioners, and where patterns in the data yield new results of interest, CDS now looms large in the future of that science.

The Crisis of Causal Knowledge

In the last few decades, as CDS was gaining in terms of the scope of the sciences it enveloped and the power of its results, philosophers such as Nancy Cartwright and philosopher/computer scientist Judea Pearl started to question whether the associations CDS found in complexes such as disease/environment and behavior/nutrition were really delivering what science ought to be delivering: robust, reproducible conclusions about causal connections in nature. In general, their worries were rather more practical than philosophical. If we want to intervene in efficacious ways to cure disease and improve human life, it would be nice to know what causes a disease—and not just what conditions (e.g., symptoms) are statistically associated with a disease state.[57]

Pearl's solution has been both a critique of the use of probabilistic reasoning through Bayesian networks—an AI technique that Pearl largely developed—and a reform program to extend the formalisms for computer-based statistical analysis to allow causal inferences to be drawn. An argument in a similar vein is presented by Nancy Cartwright, who notes that use of the associationist technique of randomized controlled trials (RCTs) does not "without a series of strong assumptions warrant predictions about what happens in practice."[58]

For Cartwright, RCTs are an important but incomplete scientific tool. In considering interventions such as giving a drug to cure a disease, they provide knowledge that the intervention "works somewhere" but fail to "clinch" the case that the same intervention will work on a different (and larger) population. This incompleteness has implications not just for the people who suffer from the disease and can be cured by the intervention—and not just for those who won't be cured by a particular intervention (and may even suffer unnecessary harm from it)—but also for large institutions like the British

[57] Judea Pearl, "Causal Inference in Statistics: An Overview," *Statistics Surveys* 3 (2009): 96–146.
[58] Nancy Cartwright, "A Philosopher's View of the Long Road from RCTs to Effectiveness," *The Lancet* 377:9775 (2011): 1400–1401.

National Health Service and other public health institutions. Interventions to cure disease cost money. Failing to cure people disappoints them.

On Cartwright's account, the difference between (statistical) association and causal knowledge is further described by a dataset and its analyses merely "vouching for" a scientific claim, as opposed to "clinching" it. Pearl echoes this call for shoring up statistical analyses: "One cannot substantiate causal claims from associations alone, even at the population level—behind every causal conclusion there must lie some causal assumption that is not testable in observational studies."[59]

These appeals for maintaining scientific reasoning with causal assumptions will sound vaguely familiar to any student of the history of modern philosophy—and indeed strikingly familiar to students of Hume's attack on causal knowledge and Kant's valiant but perhaps quixotic attempt to save us from Hume's skepticism. We can only speculate here what Hume's attitude toward CDS would have been, but given the role of the associations of ideas and impressions in Hume's epistemology and in his sentiment-associationist ethics, it seems obvious that the era of CDS would have been quite pleasing to Hume. What Hume would have found revolutionary about CDS is not only the massive amounts of data that can now be accessed (much greater than the senses, memory, and imagination can handle for a person at any one time) but also the ways in which the data can be manipulated mathematically—beyond the capabilities of the best mathematicians. Associationist knowledge in the era of CDS far exceeds the ability of one mind, and will no doubt continue to grow.

While historical questions might lead away from the primary considerations of CDS, they also serve to remind us of some of the practical restrictions that will come with pursuing the causal account of scientific knowledge. In contemporary CDS, petabytes of data are generated from millions (soon billions?) of sensors of atmospheric and terrestrial conditions. A genome from a human sample can be sequenced by a device (MinION) that plugs into a USB port on a personal computer. These examples are amazing, and there is no reason to think that the mountains of data and the power of computational techniques will not continue to increase. So where do we introduce causal assumptions to interrogate which associations are merely correlational and which are causal? CDS does not create a supermind, capable immediately of cognizing which associations are causal. Scientists will have to understand the results of CDS in order to formulate the proper causal assumptions. Causal knowledge does not come "for free."

These issues lead us to ponder what it is to have scientific understanding. David Weinberger has developed a wide-ranging critique of CDS, to the effect that it makes scientific understanding impossible for limited beings like us.[60] Studying many examples of CDS results, he concludes that:

> Clearly our computers have surpassed us in their power to discriminate, find patterns, and draw conclusions. That's one reason we use them. Rather than reducing

[59] Pearl, "Causal Inference in Statistics," 99.
[60] David Weinberger, *Too Big to Know* (New York: Basic Books, 2011).

> phenomena to fit a relatively simple model, we can now let our computers make models as big as they need to. But this also seems to mean that what we know depends upon the output of machines the functioning of which we cannot follow, explain, or understand.[61]

The lesson to take away here is that scientific understanding is a retrospective and not a time-slice activity, and that it takes more effort (in the era of CDS) than does scientific discovery. It may well be the case that CDS produces some results that boggle the mind, yet do not increase scientific understanding, after being considered "in the fullness of time." Some of these results (just like in non-CDS science) will end up being not reproducible—hence not good science. The major difference seems to be the volume of scientific results available through CDS and the speed at which these results are produced. Here the concern seems primarily practical and not epistemic in nature. That is, CDS does not seem to produce a kind of science that is in principle not understandable. So for some time it may well be wise for scientists to follow the motto of "less is more." And for any ethics of AI that is developed on the back of that science, a corresponding caution will be called for.

How an Epistemic Crisis Could Become an Ethical Crisis

Forswearing caution, some scientists have pursued CDS in publishing results of statistical correlation systems that purport to draw conclusions about people and predict their behavior. We now have techniques of whole-genome sequencing that correlate phenotypes with genomes—not merely with single or multiple genes. Christoph Lippert and his colleagues from the Venter lab discovered a technique for the "[i]dentification of individuals by trait prediction using whole-genome sequencing data," but at the same time acknowledged that their discovery "may allow the identification of individuals through genomics—an issue that implicates the privacy of genomic data," and further that their work "challenges current conceptions of genomic privacy... the adequacy of informed consent, the viability and value of deidentification of data, the potential for police profiling, and more."[62]

The ethical worry here is not so much that we will be able to pick people out of a crowd, based on a DNA sample (although that is fascinating!), but that we will be able to link genomes to phenotypic profiles. These profiles can be physiological, as in the studies Lippert et al. did on face shape, voice, age, and body-mass index, and they may eventually be used to correlate sustained tendencies toward behavior with genomes.

[61] David Weinberger, "Our Machines Now Have Knowledge We'll Never Understand," *WIRED* (April 18, 2017), accessible at: https://www.wired.com/story/our-machines-now-have-knowledge-well-never-understand/.

[62] Lippert et al., "Identification of Individuals by Trait Prediction Using Whole-Genome Sequencing Data," *PNAS* 114:38 (2017): 10166–171.

The power and perniciousness of these forms of CDS may be clearer if we relate them to worries in ethics about the treatment of individuals when statistical and aggregative techniques are used to make social choices (i.e., for the provision of health care, tax policies, and the like). Utilitarianism is one good example of a theory that relies on these aggregative techniques; John Rawls pointed out that utilitarianism tends to deny the distinctions among persons. In general, social choice procedures for large societies under "technocratic" rule have been criticized by deontological ethicists on the grounds that such procedures require measurements that aggregate over individuals, and thus treat them as indistinguishable "receptacles" of various goods. Thus CDS applied to social choice will certainly aggregate over individuals.

Will whole-genome sequencing usher in an era of technocratic management of populations? If so, this outcome of CDS may outweigh the scientific benefit that we derive from it. We should be vigilant, but also willing to accept some of the results of CDS when they are helpful in a Paretian sense ("at least one person benefits, and no one is harmed"). When trade-offs are suggested by a social choice CDS, we will have to consider carefully whether reasonable expectations (or even rights) of individuals are being violated.

Oppositional versus Systemic Approaches

We conclude by noting that most of the standard approaches to the ethics of AI—as discussed earlier—proceed as instances of applied ethics in which human rights and interests are *opposed to* an AI technology, as though humans and technologies operate somehow independently of one another. The basic idea of the oppositional approach is that AI, left unchecked, will do bad things to us. This approach can be seen in the Policy and Investment Recommendations for Trustworthy AI from the European Union (EU) High-Level Expert Group on AI.[63] They strongly recommend a "Human-Centered Approach," which suggests that there could be other possibilities, for instance a "Machine-Centered Approach."

Yet another approach would be to consider AI as a set of technologies that are embedded in a *system* of human agents, other artificial agents, laws, nonintelligent infrastructures, and social norms. That is, the ethics of AI can be seen to involve a *sociotechnical system* that has to be designed not as an isolated technical object but with attention to the social organization in which it will operate. The more we learn about AI behaviors, the better we can adapt the rest of the system to improve outcomes or, in some cases, choose not to implement an AI to take on certain functions. The main idea here is not to

[63] European Commission, "High-Level Expert Group on Artificial Intelligence" (May 2, 2019), accessible at: https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence.

require *all* of the ethics of AI to be achieved by an AI technology. Rather, the sociotechnical system can be optimized to accommodate what AI does well and what it does poorly.

It appears that there are many ethical reasons for preferring the systemic approach to the oppositional approach, partly due to the difficulties in implementing ethics in autonomous agents and partly due to the very nature of AI. After all, applications of AI are not organic entities or systems, asserting their own autonomy. Rather, they are pieces of software and devices that exist in order to improve human life. From this perspective, it would be best to design machines that help *us* to act more ethically, which means that the goal would be neither to make machines ethical by making them free moral agents nor to make machines behave ethically in conformity to moral rules. Instead, AI can help us to be wiser by making us more aware of the consequences of our actions and consequently to be more responsible when acting. To do so, it would be necessary to understand the decisions of machines, which requires that their inferences are comprehensible to us. This corresponds to the ability of the machine to provide explanations, that is, to relate their conclusions to the values that contribute to the solution they propose. It could be that many problems in machine ethics are directly related to what is often called "explainable artificial intelligence"—to the capacity to construct understandable explanations that allow humans to argue and to discuss the decisions proposed by machines that in turn may counter humans' own arguments. This approach appears to be close to ethical collective deliberations, with human and artificial agents that would collaborate in a way inspired by Jürgen Habermas's work on the ethics of communication[64] and on deliberative democracy[65].

Conclusion

Our primary message in the preceding five sections on the ethics of the ethics of AI is that progress will be made difficult by the very nature of AI, and AI problems are not likely to yield to the "common approaches" of applied ethics. But this difficulty is the very basis of our claim that there is an ethics of the ethics of AI. Progress *matters* in this domain. Artificial intelligence is here to stay, and doing the ethics of it (or *for* it) competently can help to protect important interests, save lives, and make the world a better place. Conversely, doing the ethics of AI poorly will likely yield some regrettable results, such as mistrust between ethicists and technologists and a public that is increasingly vulnerable to something they can neither understand nor avoid.

Here we can draw out the lessons from our five challenges mentioned in the preceding discussion. First, there are conceptual ambiguities that seem endemic to the ethics of AI.

[64] Jürgen Habermas, *The Theory of Communicative Action. Vol. I: Reason and the Rationalization of Society*, trans. T. McCarthy, (Boston, MA: Beacon Press, 1984).

[65] Jürgen Habermas, *Between Facts and Norms: Contributions to a Discourse Theory of Law and Democracy*, trans. W. Rehg (Cambridge, MA: MIT Press, 1996).

For ethicists (and for the general public), it can be tempting to attribute properties to AIs that they do not have. Between philosophy, computer science, AI, futurism, and science fiction, there are partly overlapping linguistic communities that use the same words with disparate concepts. In considering specific AI applications, equivocation on terms like "intelligence," "agent," and "autonomy" can quickly produce misplaced fears or unjustified optimism. This leads us to a more general observation—that ethicists of AI must guard against overestimation and underestimation of risks. When we spin fanciful stories about the "rise of the machines" and how they threaten humanity, we worry about problems that we need not face immediately or perhaps at all. When we underestimate risks, we overlook current and near-term implementations of AI in law enforcement, national security, social media, marketing, financial institutions, and elsewhere that already affect our interests and rights negatively. Still, we are confident that we can develop ethics between these two antipodes.

For most ethicists in the rationalist tradition, there remains the hope that we can design these intelligent machines to act on an ethics that we code into them—and maybe even to develop their own ethical abilities. But every approach to implementing an ethics of AI seems to have its challenges, since ethical judgments are typically defeasible, ethical behavior is difficult to model, ethical norms often conflict, and most ethical deliberations depend on judgments (i.e., discrimination) that are already difficult for humans as well as for machines. When we turn to the epistemology of the ethics of AI, we find that an ethics of AI will depend on the very science that AI produces. Unfortunately, AI plays a major role in producing scientific information without a corresponding increase in understanding. Many socially directed applications of AI will depend on scientific knowledge, but it is unclear whether humans will possess that knowledge, even though the data and analyses may advise interventions in health care, economics, environmental protection, and other areas crucial to our well-being. Finally, it will be important to reconceive the problem of the ethics of AI as a joint sociotechnical creation, and not as a series of technical problems to be confronted by better engineering. We will not be able to simply "design" away problems in the ethics of AI by controlling or opposing AI applications. We will have to see AI as a partner, of sorts, in a larger project to build better societies.

Bibliography

Arkin, Ronald C. *Governing Lethal Behavior in Autonomous Robots*. New York: Chapman and Hall/CRC Press, 2009.

Awad, Edmond, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. "The Moral Machine Experiment." *Nature* 563 (2018): 59–64.

Dennett, Daniel C. *The Intentional Stance*. Cambridge, MA: MIT Press, 1987.

Horty, John F. *Agency and Deontic Logic*. Oxford: Oxford University Press, 2001.

Kurzweil, Ray. *The Singularity Is Near: When Humans Transcend Biology*. New York: Penguin, 2006.

Lin, Patrick, Keith Abney, and Ryan Jenkins, eds. *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence*. New York: Oxford University Press, 2017.

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

Weinberger, David. *Too Big to Know*. New York: Basic Books, 2011.