Artificial Intelligence: Power to the People

Heather M. Roff

Artificial intelligence (AI) is said to hold both promise and peril. Many tout its potential for great benefits in both civilian and military applications, from greater economic growth in all sectors of the economy to more humane and precise warfare across the globe. Others, however, claim that these same civil and military applications will be increasingly harmful, causing such problems as widespread economic disruption, job loss, or even "flash wars" devoid of meaningful human control. These discussions continue into such other international realms as global health and human rights. However, much of the discussion on both sides conflates issues of automation, autonomy, AI, and machine learning, leading to a cacophony of voices speaking past and over one another with little understanding of or coherence with the realities of a coming AI future. This essay seeks to bring some calm and clarity to the discussion. If not exactly clearing the conceptual underbrush, it highlights many existing tensions that need to be resolved. Only with a clear-eyed appraisal of what AI really is can we move to addressing various normative questions about not merely its particular applications but also the ethical implications of those applications in the international system.

This essay proceeds in three sections. The first suggests we need to be careful of conflating AI with automation or autonomy, for doing so risks aggregating benefits and harms in different ways, when we would do better to keep them separate. The second section argues that AI can be applied in all sectors of the economy as well as in warfare but that to understand the applications for which AI is best suited, we must have a clear understanding of the problems that AI is best prepared to address. All problems are not equal, and not all problems can be solved with algorithmic estimations. I submit that much of the AI landscape revolves

Ethics & International Affairs, 33, no. 2 (2019), pp. 127–140.

doi:10.1017/S0892679419000121

around epistemological questions that are not, in fact, particular to AI. Finally, I conclude that when it comes to normative considerations, AI is not and will never be a "moral agent," and thus cannot determine what is morally right. Indeed, I suggest that AI presents us with a classic manifestation of Hume's guillotine: We cannot speak about ethical AI because all AI is based on empirical observations; we cannot get an "ought" from an "is." If we are clear eyed about how we build, design, and deploy AI, we will conclude that all of the normative questions surrounding its development and deployment are those that humans have posed for millennia.

Fuzzy Logic? Defining Automation, Autonomy, and AI

It is quite common to find discussions about automation, autonomy, and AI happening within the same breath, with policymakers, scholars, and practitioners at times seeming to use them interchangeably. These terms are not, in fact, interchangeable, and each of them has a particular meaning. As we begin to grapple with difficult normative and policy questions related to such new technology, it is essential to understand what each implies, as well as what each does not.

Automation

Though the term "automation" was not coined until 1946 by Ford Motor Company engineer D. S. Harder, at a basic level automation has a much longer history when viewed from the perspective of mechanizing labor. From the printing press to early conceptions of the computer, such as Countess Ada Lovelace's analytical engine, to its more contemporary usage in the early twentieth century by automobile manufacturers, automation has existed in myriad forms for hundreds of years. One early paper on automation and industrial safety written in 1958 proposed that "automation itself begins only when you combine several similar operations mechanically and add the necessary controls to provide continuous automatic production until the product is completed."[1] While this definition aims to highlight the importance of process and control, it is rather dated in viewing automation as needing to be directed toward production; however, it still seems quite appropriate if we exchange "product" for more contemporary terms like "output." Regardless, clearly automation is not in itself intelligent. It is merely the ability of a system to carry out a task—without deviation—automatically.[2] The machine is on; it runs its course. Take, for example, the first steam-powered vehicles. From simple wagons to trains, these systems relied on pistons, cranks,

https://doi.org/10.1017/S0892679419000121 Published online by Cambridge University Press

wheels, and steam to mechanize the transport of everything from grain to people. Such systems can be more or less complex in their functions and tasks, but no AI is required.

Autonomy

Today we are presented with a new puzzle: confusion over the difference between automated and autonomous systems. This confusion may be due in part to the greater reliance on information-processing systems in all aspects of our lives, unfamiliarity with how such systems operate, and circular definitions. How does autonomy differ from automation?

Some claim that autonomy is merely the ability of a system to act without the intervention of an operator.[3] On this minimal approach to defining autonomy, there does not appear to be a bright line between automated and autonomous systems. After all, automated systems also act without intervention of a human operator. Attempts to square this circle take a "functional approach," decomposing a system into particular tasks or functions. Tasks or functions can be coupled in various ways depending on the system, and they can also be of higher or lower difficulty. For example, many cars have some functions that operate without any input from the driver, such as auto-park. In this instance, the minimal approach would say that the car has certain autonomous functions but that the overall car is not "fully" autonomous. Here, a human performs some tasks but not others, and so there are differing degrees of human-machine interaction. Nevertheless, at its core this minimal definition seems to preclude any meaningful difference between automation and autonomy because both rely on the absence of human intervention.[4]

Others have offered more robust definitions of "autonomous," including that "a system must have the capability to independently compose and select among different courses of action to accomplish goals based on its knowledge and understanding of the world, itself, and the situation."[5] This approach suggests that it is the *entirety* of the system, not just the independent tasks within the system, that needs to function without intervention. Further, the system must also have higher-order cognitive capacities for more complex forms of reasoning. In other words, this more robust definition implies that the system must have some form of *intelligence*.

The concept of "intelligence," however, is a slippery one. How we define, and eventually measure, intelligence in natural beings such as humans is notoriously

https://doi.org/10.1017/S0892679419000121 Published online by Cambridge University Press

difficult. Whether there are different forms of intelligence such as, for example, theoretical knowledge and emotional understanding (IQ versus EQ) is still debated today. And, as Stephen Cave has convincingly argued, throughout history those in power have gotten to decide what counts as intelligence, changing its definition to include and exclude different peoples, often using it as a tool of domination.[6] As Cave suggests, the answer to the question of who is and who is not intelligent has serious social and political implications, such as who is accorded privileges, status, and rights.

Despite these challenges, the simple fact is that we must find some basic definition if we are to have clarity. For our purposes, we can at least claim that there is some agreement among psychology, computer science, and engineering that intelligence, prima facie, requires having some sort of purposiveness and the requisite faculties to pursue a goal or end.[7] These faculties could include sensing, perception, reasoning, and a means by which to act—say, through physical embodiment.

Artificial Intelligence?

However, even if we "agree" on what to include in the means-ends relationship of intelligence, we are still left with some serious conceptual difficulties when we begin speaking about "artificial intelligence."[8] I suggest that these difficulties arise not only from pure definitional disagreements but from the tendency (outlined above) of scholars and practitioners to move the goal posts on ontological questions about what constitutes intelligence and, thus, what constitutes AI. For instance, some claim AI is "the field devoted to building artificial animals (or at least artificial creatures that—in suitable contexts—*appear* to be animals) and, for many, artificial persons (or at least artificial creatures that—in suitable contexts—*appear* to be persons)."[9] Even more generally, Margaret Boden claims that AI "seeks to make computers do the sorts of things minds can do."[10] In other considerations, AI "is a cross-disciplinary approach to understanding, modeling, and replicating intelligence and cognitive processes by invoking various computational, mathematical, logical, mechanical, and even biological principles and devices."[11] Or, as Stuart Russell claims, AI is nothing more than building intelligent agents, where these agents are said to possess "rationality."[12]

Each of these definitions proves problematic in one form or another, and thus leaves confusion in its wake. Boden's more general treatment seems straightforward until one inquires as to what constitutes a "mind," which can lead to endless debates. Relying on Russell's notion of rationality would seem to preclude

https://doi.org/10.1017/S0892679419000121 Published online by Cambridge University Press

approaches to broadening the set of possible intelligences to include non-human animals, not to mention that it would unjustifiably privilege classic Enlightenment conceptions of human beings. Finally, defining AI as an “approach to understanding . . . intelligence” kicks the proverbial can down the road about what exactly constitutes intelligence in the first place and what makes *artificial* intelligence different than its natural forms. We are still left without a clear understanding of what AI actually is and which systems can be considered artificially intelligent.

Let us examine a seemingly simple example. Say we build a symbolic logic expert system—essentially an intricate decision tree matrix with hand-coded knowledge and rules. One modern example of this would be something like the popular tax software programs that help us file our taxes. A human user inputs information such as whether she is married or single, and the system has rules for how to respond depending on that input. Such a system is what John Haugeland dubbed “GOFAI,” or “good old-fashioned AI,”[13] and we may want to agree that this system is artificially intelligent. However, what allows this system to be classified as AI is the knowledge base that is preprogrammed into it. The software, for example, does not understand what being married actually means; it only knows the certain rules it has to follow if someone clicks the “married” option. Moreover, the computational architecture does not permit deviations; it only allows a strictly limited set of actions. So when the tax code changes, a human must rewrite the software accordingly; the system cannot change the rules on its own. In short, symbolic AI systems do not *learn*, and they do not *adapt*.

However, if one claims that this preprogrammed logic, even if highly complex, is not *really* AI, and that only systems that can adapt and learn are truly intelligent, then this puts us in a different sort of position—for three reasons. First, we would have to concede that many forms of automation in today’s societies have nothing to do with AI, despite the history of the field. If this is so, then many of the worries and fears about the coming AI boom and doom are misplaced. We could not say that AI is putting people out of work or changing the distribution of wealth and the character of warfare; we would have to acknowledge that it is merely a type of automation running on computational control processes. If, however, one were to claim that GOFAI systems are AI systems, then we would need to find some way of parsing how the “future” of AI, with all its benefits and risks, is something

https://doi.org/10.1017/S0892679419000121 Published online by Cambridge University Press

different than what has come before. We would need an explanation of why AI is now something to either welcome or fear.

Second, we would have to argue that non–logic-based techniques for machine learning—where instead of having preprogrammed rules, an algorithm generates its own rules based on training data that (for now) a human has fed it—become the true signifier for AI.[14] However, there is no agreement that this is the case. While machine learning, and deep learning in particular, has gained much ground and public attention over the past fifteen years, these techniques are ultimately still only narrowly applied to particular tasks. Indeed, such machine-learning AI agents can learn—by themselves—the underlying structures of provided data, but this is only to say that they can find patterns in the data, even when those patterns include empirically or normatively spurious correlations. Their intelligence is limited to the domain and task for which they are trained, and they (currently) remain incapable of learning more robust or general concepts.

These systems come in a variety of forms. Some systems may be able to continually learn and adapt through the intake of new data (known as "online learning" or "lifelong learning"), but these systems are easily manipulated in a variety of ways.[15] Other systems are not continuously learning but trained on labeled data and "frozen" before deployment. In both instances we have few testing, verification, and validation methods to understand exactly how the systems do what they do, and whether they will continue to perform as intended. For these reasons, much attention is now being paid to "explainable AI" as a way of having the system tell a human how it came to its conclusion with some uncertainty estimation.[16] Regardless, if AI is simply synonymous with machine learning, then we must limit the discussion of AI's benefits and harms to only machine-learning systems. We must also then discern whether those considerations are truly different from those surrounding mere automation. One possibility is that simply employing these types of systems en masse and at scale to a much wider variety of tasks than ever before will result in truly new and different concerns. However, the jury is still out.

Third, depending upon which version of "autonomy" we use, machine learning may have little to do with AI. For example, if we take autonomy to simply mean there is no human in the loop, then we may want to resist tendencies to run the two together. Like a Venn diagram, they may be very different but overlap in some instances, but it depends on where we draw the lines. Moreover, even if we accept a broader definition of autonomy, autonomous systems may or may not be

https://doi.org/10.1017/S0892679419000121 Published online by Cambridge University Press

lifelong learners, for even the broadest definitions do not seem to require anything like this capability.

The point is that we have little conceptual clarity in the great AI debate because there is little agreement on what to include or exclude. If we cannot find the lines of inclusion and exclusion, then having any serious discussion about the possible benefits and risks of AI is almost impossible. Indeed, even discussing "AI governance," including hard law, professional standards, and norms, presupposes having a finer-grained level of detail than currently exists, despite the many approaches proffered to date.[17]

AI for All?

It should become apparent that however we decide to define or to limit discussion of what AI is, such as whether it includes GOFAI or is merely machine learning, AI can be applied almost anywhere. After all, to create and tailor AI for a given situation, we only need four things: computing power, expertise, data, and an algorithm. Thus, as society increasingly produces and sheds data, and the capabilities of computer processing and transfer speeds increase, AI will undoubtedly be applied simply because *it can be*. Most of the necessary algorithms are decades old, open source, and easily accessed. Expertise, however, remains the key to unlocking AI's potential, placing computer engineers—as well as those with technological prowess in the policy, civil society, and academic realms—in particular positions of power.[18]

Despite the shifting power balance, there are some ways to democratize these powers even more. This is due to some simple facts about the potential scale and use of AI. In other words, for a problem to be well suited for an AI application, we really need to have a lot of latent and background understanding about that problem, something that is often not appreciated. To explain, there are some basic notions from control theory that underpin most AI systems. Control theory deals with dynamic systems; that is, systems that interact with their environment. We might control a system by changing the desired task, the required sensors, the data gathered, the information processes, or the guidance or architecture for how the system acts or reacts. The actions of the system can be few or many, and the choice of what to pursue can be narrow or wide. But what its construction ultimately requires is a robust understanding of what the system needs to function *appropriately*.

https://doi.org/10.1017/S0892679419000121 Published online by Cambridge University Press

"Appropriate," here, is the crucial word. For we must not only be able to clearly identify the desired task and construct a system that is capable of carrying it out but also be able to understand how and in what manner the task was completed. This is where the boundary between development and deployment matters most, because there are all sorts of other metrics that come into play. For instance, did the system act safely? Did it act in accordance with laws, regulations, or norms? Answering these questions allows us to gain a clearer understanding of which problems are well suited for AI and which are more troublesome. Indeed, it also allows us to acknowledge that some problems cannot be solved with AI.

For tasks that are well defined and narrow, such as calculating the estimated trajectories of an object, an AI agent can do quite well. This is because we understand how gravity, drag, velocity, and the general laws of physics work. For such problems, even if we do not tell an agent beforehand what these properties are but allow it to learn them through iteration, given sufficient interaction with the environment, it can figure out how to do something even if not understanding why.[19] But when it turns to more general tasks, an AI agent finds much more difficulty. The context of each skill or task may change, the environment may change, and the reasoning required to accomplish different tasks may be of a much higher order. This is mostly due to the fact that transferring how an agent perceives, reasons about, or finds patterns within its input data to another domain or task is extremely difficult, often causing the entire system to fail.[20]

It is no surprise, then, that many of the areas where AI is seen as being most beneficial are areas with large amounts of data to draw on, where the structure of the task is generally known, and where there is a way to easily check whether it has correctly and appropriately completed its task. For example, limited forms of medical diagnostics and natural language processing have both seen major advancements thanks to AI. However, many of the applications that scholars and policymakers have proposed for AI are embedded in social and political structures, and the decisions and actions that an AI system takes will affect those structures. For example, many in the West decry the Chinese government's use of a social credit system to monitor and limit the actions of its people. If one believes that such surveillance and the limitation of particular freedoms is a violation of human rights, then that person might argue that the mere availability of AI to enable those rights violations could further entrench beliefs about particular classes of people, could give rise to new vulnerable populations, and could even lead to destabilizing conditions in the international system.[21]

https://doi.org/10.1017/S0892679419000121 Published online by Cambridge University Press

An Epistemological Word of Caution

As I have hinted at, many of the questions in the AI debate are actually epistemological questions, and they are not confined to AI. Rather, they are age-old questions about whether we have some ground truths from which even to proceed. Epistemology, or a theory of knowledge, for an AI is necessary because the notion of an AI is premised on the derivative concept of intelligence. A theory of knowledge, with its attendant requirement of offering differing types of knowledge and standards for justified belief, validity, error, and the like, is required if we are to make claims about whether the system is acting intelligently and appropriately. Additionally, what constitutes knowledge for an AI is tightly coupled with such things as what sorts of information it can access, identify, and assess; the formal relation between its capacity to reason, its sensory abilities, and the decisions it makes; and its ability to update and adapt based on interactions with its environment, other agents (human users or other AIs), or even simulated data.

Such epistemological grounding is further important because in cases where an agent may be acting in an environment that it perceives incorrectly, it may "believe" that it is acting correctly according to some objective reality, but in fact, through various potential errors, this belief is erroneous.[22] Thus, when we design such systems, we must take due care in ensuring that the task, skill, or goal given to an AI is one that it can adequately measure against an objective reality. If designers fail to adequately match the AI's goals to the correct measurements and variables, then the agent's perceived and/or learned knowledge will not correspond to reality. We can call this "knowledge correspondence." A well-designed AI system will have 100 percent knowledge correspondence. As the distance increases between the agent's perception and the agreed-upon ground truth, error and bias will emerge. The lower the correspondence, the greater the error. But to once again reiterate: If we are to build systems that are capable of 100 percent knowledge correspondence, then we first must understand what objective reality is or whether it even exists.

This may seem like a truism but, in effect, we must acknowledge that humans live in a complex web of social interactions, norms, customs, and power relations. In this web, there are multiple "subjectivities" at play. There may be some form of intersubjective objectivity—that is, we all subjectively agree on what constitutes reality and truth—but the basis of this is purely a subjective feeling, experience,

https://doi.org/10.1017/S0892679419000121 Published online by Cambridge University Press

or thought. Unlike the laws of physics, social "realities" are constructed, subject to change, and not universal.

Why does this matter for understanding the future of AI? It matters on two fronts. The first is related to whether we can, from a technological standpoint, create AI systems to carry out tasks related to the subjectivities noted above.[23] That is, a developer might ask, do I possess sufficient knowledge about a certain behavior, norm, or intersubjective practice to create and train an agent? Here, we are presented with questions about the type and quality of data given to train an agent. For instance, I may be able to train an autonomous vehicle to drive on the left versus the right side of the road, but I may not be able to train an AI to identify who is a combatant in war. I may be able to train an AI agent to identify certain types of weapons or even people, but I currently cannot train it to understand the more fluid and dynamic notion of combatancy. This is because combatancy is a behavior that is dependent upon hostile *intent*, and there are myriad forms that it may take in a variety of environments, with almost infinite kinds of weapons. Thus, when discussing the morality or potential efficacy of autonomous weapons systems, one must first determine whether the system is capable of attacking particular kinds of targets and how it came to those determinations.

On the second front, when we talk about norms or social concepts, we also face the question of "operationalization"—that is, how we move from a concept in theory to real decision-making action on the ground. Operationalizing a social concept well is not an easy task. Which variables one identifies and which proxies one chooses to represent the concept are both in some way value loaded. This is because we are dealing with concepts such as justice, equality, and fairness. Depending upon where one sits, one may choose different values to weigh. Or one may even object to the entire notion of "weighing" for claims of justice.

These are no small matters. If one believes that crime is something that ought to be avoided because it harms individuals and societies, then that person might think that designing a system that can identify those most likely to commit or recommit crimes is a good application of AI.[24] However, if, in doing so, the person does not consider the wider structural injustices that may be at play in gathering the data to train these systems, he or she will not produce a useful or morally good use of AI but merely mirror the existing structural input biases in the outputs.

https://doi.org/10.1017/S0892679419000121 Published online by Cambridge University Press

Next Steps

There are two areas that need immediate attention if we are to realize the benefits and minimize the harms of AI. First, on the technological side, we must acknowledge there is no overarching theory or principle by which computer scientists decide how to weigh values in their systems. Second, on the empirical and normative side, we must have discussions from an interdisciplinary perspective about what types of knowledge are appropriate for various uses of AI. In essence, if we want to use AI in applications that deal with tasks concerning human behavior and that affect rights, then we first need good theories about how people do and ought to act. Building such theories requires us to think deeply about what the objective reality is, whether there is one, whether we can measure it, and whether this reality is even something we want. After all, that reality may be unjust, unfair, or harmful. But these questions are not questions of AI. These are the deep and difficult questions of ethics.

The Impossibility of a Moral Machine?

Math does not morals make. This is a simple but effective truth in the great AI debate. An AI can never be a moral agent, even if it can reason in some way or can choose to act from among ways X, Y, and Z. Granted, algorithms and data are effectively complex adaptive systems capable of accomplishing great and interesting things that are at times very useful and at times very worrying. But the AI agent is not a moral agent. It is we humans who control it, and we who define whether it is functioning appropriately. Therefore, it is not in the position to be accorded rights and duties.

Moreover, no matter how sophisticated an AI agent becomes in its reasoning, one must grant that at a base level it is a mathematical model finding correlations in data too huge for humans to assess. It may find novel "insights," and it may be "creative" in its solutions. Indeed, those insights may disrupt our current ways of understanding or doing things. But the simple fact remains that any AI system is built on the data it consumes. The data it consumes comes from empirical facts about how it observes the world.

If we believe that an AI system will thus generate a morally right answer, yield a morally correct course of action, or determine who a morally good person is, then we have walked straight into Hume's guillotine, assuming that a machine can derive an "ought" from an "is." For an AI system cannot learn without empirical

https://doi.org/10.1017/S0892679419000121 Published online by Cambridge University Press

data. Even simulated data is built on real data. If we universally accept, as Hume suggests, that one cannot yield a normative judgment simply from a set of empirical physical facts, then AI systems will never be morally prescriptive.[25] We cannot grant one without cutting off our head for the other.

Rather, if we want AI systems to carry out their tasks ethically, then we need to train them to do so. We cannot merely feed in massive amounts of data and believe that the tyranny of the masses will not result. Instead, we must be clear eyed about AI's capabilities, its limitations, what problems it is best suited to solve, and which ones require insights and latent knowledge often not found in the computer science discipline. Only agents capable of being ethical can design systems in a thoughtful and responsible way; only they are moral agents.

NOTES

[1] Donald F. Harrington, "Automation's Impact on Industrial Safety," *Cleveland-Marshall Law Review* 7, no. 2 (1958), p. 266.

[2] Defense Science Board, *Summer Study on Autonomy* (Washington, D.C.: Office of the Under Secretary of Defense for Acquisition, Technology and Logisitics, June 2016), p. 4, www.hsdl.org/?view&did=794641.

[3] "Autonomy in Weapons Systems" (DoD Directive 3000.09, United States Department of Defense, November 21, 2012), www.esd.whs.mil/portals/54/documents/dd/issuances/dodd/300009p.pdf. See also United States Office of the Director of National Intelligence, *The Aim Initiative: A Strategy for Augmenting Intelligence Using Machines* (Washington, D.C.: Office of the Director of National Intelligence, 2019), www.dni.gov/index.php/newsroom/reports-publications/item/1940-the-aim-initiative-a-strategy-for-augmenting-intelligence-using-machines.

[4] For instance, the National Highway Traffic Safety Administration (NHTSA) in the United States demarcates a scale of 0–5, according to which "0" indicates a vehicle where all tasks are performed by a human driver and "5" indicates a vehicle that is capable of all tasks related to driving but for which the human driver may intervene and take control. Indeed, the NHTSA actually uses the two concepts interchangeably in its description of level 0, noting "zero autonomy" in the explanation but "no automation" in the header. See "Automated Vehicles for Safety," NHTSA, www.nhtsa.gov/technology-innovation/automated-vehicles-safety.

[5] Defense Science Board, *Summer Study on Autonomy*, p. 4.

[6] Stephen Cave, "Intelligence: A History," *Aeon*, February 21, 2017, aeon.co/essays/on-the-dark-history-of-intelligence-as-domination.

[7] Maja J. Mataric, "Designing Emergent Behaviors: From Local Interactions to Collective Intelligence," in Jean-Arcady Meyer, Herbert L. Roitblat, and Stewart W. Wilson, eds., *From Animals to Animats 2: Proceedings of the Second International Conference on Simulation of Adaptive Behavior* (Cambridge, Mass.: MIT Press, 1993), pp. 432–40; Stuart J. Russell and Peter Norvig, *Artificial Intelligence: A Modern Approach* (Upper Saddle River, N.J.: Prentice Hall, 1995); Robert J. Sternberg, *Beyond IQ: A Triarchic Theory of Human Intelligence* (Cambridge, U.K.: Cambridge University Press, 1985); C. F. Schmidt, N. S. Sridharan, and J. L. Goodson, "The Plan Recognition Problem: An Intersection of Psychology and Artificial Intelligence," *Artificial Intelligence* 11, nos. 1–2 (August 1978), pp. 45–83; Edward Chace Tolman, *Purposive Behavior in Animals and Men* (Berkeley: University of California Press, 1949); Michael van Lent, John Laird, Josh Buckman, Joe Hartford, Steve Houchard, Kurt Steinkraus, and Russ Tedrake, "Intelligent Agents in Computer Games," in *Proceedings of the National Conference on Artificial Intelligence* 16 (1999), www.aaai.org/Papers/AAAI/1999/AAAI99-143.pdf.

[8] Max Tegmark claims that "intelligence is the ability of an entity to achieve complex goals." In some senses this is correct. However, this definition is overly general and relies heavily on tacit knowledge about the kinds of faculties an agent needs to possess to have this ability, as well as being overly specific in the complexity of the goals sought. Some may have very simple goals, but the routes by which they

https://doi.org/10.1017/S0892679419000121 Published online by Cambridge University Press

attain them may themselves be highly complex. See Max Tegmark, *Life 3:0: Being Human in the Age of Artificial Intelligence* (New York: Alfred A. Knopf, 2017), p. 280.

[9] Selmer Bringsjord and Naveen Sundar Govindarajulu, "Artificial Intelligence," in *Stanford Encyclopedia of Philosophy*, 2018, plato.stanford.edu/entries/artificial-intelligence/ (italics in original).

[10] Margaret A. Boden, *AI: Its Nature and Future* (New York: Oxford University Press, 2016), p. 1.

[11] Keith Frankish and William M. Ramsey, introduction to *The Cambridge Handbook of Artificial Intelligence*, edited by Keith Frankish and William M. Ramsey (Cambridge, U.K.: Cambridge University Press, 2014), p. 1.

[12] Stuart J. Russell, "Rationality and Intelligence," *Artificial Intelligence* 94, nos. 1–2 (July 1997), pp. 57–77.

[13] John Haugeland, *Artificial Intelligence: The Very Idea* (Cambridge, Mass.: MIT Press, 1985).

[14] These techniques would include convolutional neural networks, deep neural networks, reinforcement learning, general adversarial networks, and the like.

[15] See, for example, much of Jeff Clune's work on adversarial examples and machine vision. There is also concern for model inversion or data-poisoning attacks. Anh Nguyen, Jason Yosinski, and Jeff Clune, "Deep Neural Networks Are Easily Fooled: High Confidence Predictions for Unrecognizable Images," in *Computer Vision and Pattern Recognition* (Los Alamitos, Calif.: IEEE Computer Society, June 2015), www.evolvingai.org/fooling.

[16] See the Defense Advanced Research Projects Agency's web page on explainable AI at David Gunning, "Explainable Artificial Intelligence (XAI)," www.darpa.mil/program/explainable-artificial-intelligence.

[17] For example, see Association for Computing Machinery Public Policy Council, "Statement on Algorithmic Transparency and Accountability," Association for Computing Machinery, January 12, 2017, www.acm.org/binaries/content/assets/public-policy/2017_usacm_statement_algorithms.pdf; "Asilomar AI Principles," Future of Life Institute, futureoflife.org/ai-principles/; *Perspectives on Issues in AI Governance*, Google AI, ai.google/static/documents/perspectives-on-issues-in-ai-governance.pdf; "About the Japanese Society for Artificial Intelligence Ethical Guidelines," AI Committee Ethics Committee, May 3, 2017, ai-elsi.org/archives/514; Great Britain, Parliament, House of Lords, Select Committee on Artificial Intelligence, *AI in the UK: Ready, Willing, and Able?* (London: House of Lords, 2018); European Commission's High-Level Expert Group on Artificial Intelligence, *Draft Ethics Guidelines for Trustworthy AI*, European Commission, December 18, 2018, ec.europa.eu/digital-single-market/en/news/draft-ethics-guidelines-trustworthy-ai; European Group on Ethics in Science and New Technologies, *Statement on Artificial Intelligence, Robotics and 'Autonomous' Systems*, European Commission, March 9, 2018, ec.europa.eu/research/ege/pdf/ege_ai_statement_2018.pdf; IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems, *Ethically Aligned Design: A Vision for Prioritizing Human Well-Being with Autonomous and Intelligent Systems*, version 2, IEEE, 2017, standards. ieee.org/develop/indconn/ec/autonomous_systems.html; *Montreal Declaration for a Responsible Development of Artificial Intelligence 2018*, Montréal Declaration Responsible AI, www.montrealdeclaration-responsibleai.com/the-declaration.

[18] Of course, expertise is a form of authority. However, I am referring here to more traditional forms of authority-related role responsibilities. Indeed, recent efforts by Google and Microsoft employees to rebuke their companies and leadership for working on projects for the U.S. military are a case in point. Google was forced to withdraw future efforts on Project Maven, a project that uses AI for video image processing from data obtained by remotely piloted aircraft. Microsoft employees likewise urged the company not to pursue a lucrative cloud-computing contract, Project JEDI, with the Pentagon because they deemed it to be unethical to aid the U.S. military in engaging in their activities. Google was also seen to be forced by its employees to withdraw from bidding on the same contract. See Scott Shane and Daisuke Wakabayashi, "'The Business of War': Google Employees Protest Work for the Pentagon," *New York Times*, April 4, 2018, www.nytimes.com/2018/04/04/technology/google-letter-ceo-pentagon-project.html; Employees of Microsoft, "An Open Letter to Microsoft: Don't Bid on the US Military's Project JEDI," Medium, October 12, 2018, medium.com/s/story/an-open-letter-to-microsoft-dont-bid-on-the-us-military-s-project-jedi-7279338b713.

[19] Nevertheless, even seemingly narrow tasks can be tricky, and another example may be useful here. In 2016, a nonprofit research organization called Open AI used reinforcement learning to train an agent to play the computer game Coast Runners. The humans presumed that the goal of the game was to score the most points to win. In training the agent, they used "points" as the reward function—the signal that tells the agent it is learning the right thing. After learning how to generally play the game, the agent began to act in an unpredictable way. It found a bug in the game and discovered that by doing donuts in one part of the course, it could indefinitely score points, but that in so doing it would never actually finish the game and would lose. This is known as a "reward hack." The narrow task of playing the game was still difficult because the reward function in this case could have been misspecified or

https://doi.org/10.1017/S0892679419000121 Published online by Cambridge University Press

underspecified, or it could have just been a flaw in the environment itself that the agent learned to exploit, much to the chagrin of its developers. See Dario Amodei and Jack Clark, "Faulty Reward Functions in the Wild," OpenAI, December 21, 2016, blog.openai.com/faulty-reward-functions/.

[20] This is known as a problem of "catastrophic forgetting."

[21] Joe McDonald, "China Bars Millions from Travel for 'Social Credit' Offenses," AP, February 22, 2019, www.apnews.com/9d43f4b74260411797043ddd391c13d8.

[22] "Agent" here refers to both a human agent and an AI agent. One is obviously much more philosophically rich, but for AI an agent is considered one that is merely capable of observing its environment and acting on or in that environment to achieve a goal. See Russell and Norvig, *Artificial Intelligence*, ch. 2.

[23] For example, Yann LeCun, the vice president and chief AI scientist at Facebook and Silver Professor of Dara Science, Computer Science, Neural Science, and Electrical Engineering at New York University, has recently stated that AI needs more "theory building" from mathematics. This is somewhat true but misses the point that every discipline has multiple theories on which to draw, particularly in relation to human behavior. See Yann LeCun (@ylecun), "Deep learning needs more theory. Many workshops I have helped organize in the last decade . . . ," Twitter post, February 2, 2019, twitter.com/ylecun/status/1091732463284563968.

[24] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner, "Machine Bias: There's Software Used across the Country to Predict Future Criminals. And It's Biased against Blacks," ProPublica, May 23, 2016, www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[25] A recent paper attempts to argue against this conclusion by positing an "anchor" to "connect behavior to intrinsic or nonderivative values" for an AI agent. This anchor presumably allows the agent to align its behavior to that of some normative standard. However, the authors' argument falls squarely back into Hume's guillotine because, by training the agent on empirical cases, it cannot arrive at a normative conclusion. See Tae Wan Kim, Thomas Donaldson, and John Hooker, *Mimetic vs Anchored Value Alignment in Artificial Intelligence*, October 25, 2018, arxiv.org/pdf/1810.11116.pdf.

Abstract: To adequately estimate the beneficial and harmful effects of artificial intelligence (AI), we must first have a clear understanding of what AI is and what it is not. We need to draw important conceptual and definitional boundaries to ensure we accurately estimate and measure the impacts of AI from both empirical and normative standpoints. This essay argues that we should not conflate AI with automation or autonomy but keep them conceptually separate. Moreover, it suggests that once we have a broad understanding of what constitutes AI, we will see that it can be applied to all sectors of the economy and in warfare. However, it cautions that we must be careful where we apply AI, for in some cases there are serious epistemological concerns about whether we have an appropriate level of knowledge to create such systems. Opening the aperture to include such questions allows us to further see that while AI systems will be deployed in a myriad of forms, with greater or lesser cognitive abilities, these systems ought never to be considered moral agents. They cannot possess rights, and they do not have any duties.

Keywords: artificial intelligence, machine learning, autonomy, automation, epistemology, ethics

https://doi.org/10.1017/S0892679419000121 Published online by Cambridge University Press