CHAPTER 1

THE ARTIFICIAL INTELLIGENCE OF THE ETHICS OF ARTIFICIAL INTELLIGENCE

An Introductory Overview for Law and Regulation

JOANNA J. BRYSON

For many decades, artificial intelligence (AI) has been a schizophrenic field pursuing two different goals: an improved understanding of computer science through the use of the psychological sciences; and an improved understanding of the psychological sciences through the use of computer science. Although apparently orthogonal, these goals have been seen as complementary since progress on one often informs or even advances the other. Indeed, we have found two factors that have proven to unify the two pursuits. First, the costs of computation and indeed what is actually computable are facts of nature that constrain both natural and artificial intelligence. Second, given the constraints of computability and the costs of computation, greater intelligence relies on the reuse of prior computation. Therefore, to the extent that both natural and artificial intelligence are able to reuse the findings of prior computation, both pursuits can be advanced at once.

Neither of the dual pursuits of AI entirely readied researchers for the now glaringly evident ethical importance of the field. Intelligence is a key component of nearly every human social endeavor, and our social endeavors constitute most activities for which we have explicit, conscious awareness. Social endeavors are also the purview of law and, more generally, of politics and diplomacy. In short, everything humans deliberately do has been altered by the digital revolution, as well as much of what we do unthinkingly.

Often this alteration is in terms of how we can do what we do—for example, how we check the spelling of a document; book travel; recall when we last contacted a particular employee, client, or politician; plan our budgets; influence voters from other countries; decide what movie to watch; earn money from performing artistically; discover sexual or life partners; and so on. But what makes the impact ubiquitous is that everything we have done, or chosen not to do, is at least in theory knowable. This awareness fundamentally alters our society because it alters not only how we can act directly, but also how and how well we can know and regulate ourselves and each other.

A great deal has been written about AI ethics recently. But unfortunately many of these discussions have not focused either on the science of what is computable or on the social science of how ready access to more information and more (but mechanical) computational power has altered human lives and behavior. Rather, a great deal of these studies focus on AI as a thought experiment or "intuition pump" through which we can better understand the human condition or the nature of ethical obligation. In this *Handbook*, the focus is on the law—the day-to-day means by which we regulate our societies and defend our liberties. This chapter sets out the context for the volume by introducing AI as an applied discipline of science and engineering.

Intelligence Is an Ordinary Process

For the purpose of this introduction, I will use an exceedingly well-established definition of intelligence, dating to a seminal monograph on animal behavior.[1] *Intelligence* is the capacity to do the right thing at the right time. It is the ability to respond to the opportunities and challenges presented by a context. This simple definition is important because it demystifies intelligence, and through it AI. It clarifies both intelligence's limits and our own social responsibilities in two ways.

First, note that intelligence is a process, one that operates at a place and in a moment. It is a special case of *computation*, which is the physical transformation of information.[2] Information is not an abstraction.[3] It is physically manifested in energy (light or sound), or materials. Computation and intelligence are therefore also not abstractions. They require time, space, and energy. This is why—when you get down to it—no one is really ever that smart. It is physically impossible to think of everything. We can make trade-offs: we can, for example, double the number of computers we use and cut the time of a computation nearly in half. The time is never cut quite in half, because there is always an

[1] George John Romanes, *Animal Intelligence* (London: D. Appleton, 1882).

[2] Michael Sipser, *Introduction to the Theory of Computation*, 2nd ed. (Boston: PWS, Thompson, 2005).

[3] Claude Elwood Shannon, "A Mathematical Theory of Communication," in *Bell System Tech. J.* 27.3 (1948): 379–423.

extra cost of splitting the task and recombining the outcomes of the processing.[4] But this near halving requires fully double the space for our two computers, and double the energy in the moment of computation. The sum of the total energy used is again slightly more than the same as for the original single computer, due again to extra energy needed for the overheads. There is no evidence that quantum computing will change this cost equation fundamentally: it should save not only on time but also on space, however the energy costs are poorly understood and to date look fiendishly high.

Second, note that the difference between *intelligence* and *artificial intelligence* is only a qualifier. *Artificial* means that something has been made through a human process. This means by default that humans are responsible for it. The artifact actually even more interesting than AI here is a concept: *responsible*. Other animals can be trained to intentionally limit where they place (for example) even the fairly unintentional byproducts of their digestive process, but as far as we know only humans have, can communicate about, and—crucially—can negotiate an explicit concept of responsibility.

Over time, as we recognize more consequences of our actions, our societies tend to give us both responsibility and accountability for these consequences—credit and blame depending on whether the consequences are positive or negative. Artificial intelligence only changes our responsibility as a special case of changing every other part of our social behavior. Digital technology provides us with *better* capacity to perceive and maintain accounts of actions and consequences, so it should be easier, not harder, to maintain responsibility and enforce the law. However, whether accountability is easier with AI depends on whether and in what ways we deploy the capacities digital technology affords. Without care and proper measures, the increased capacity for communication that information communication technology (ICT) provides may be used to diffuse or obscure responsibility. One solution is to recognize the lack of such care and measures for promoting accountability in processes concerning digital artifacts to be a form of negligence under the law. Similarly, we could declare that unnecessary obfuscation of public or commercial processes is a deliberate and culpable evasion of responsibility.

Note that the simplicity of the definitions introduced in this section is extremely important as we move toward law and regulation of systems and societies infused with AI. In order to evade regulation or responsibility, the definition of intelligence is often complicated in manifestos by notions such as sentience, consciousness, intentionality, and so forth. I will return to these issues later in the chapter, but what is essential when considering AI in the context of law is the understanding that no fact of either biology (the study of life) or computer science (the study of what is computable) names a necessary point at which human responsibility should end. Responsibility is not a fact of nature. Rather, the problem of governance is as always to design our artifacts—including the law itself—in a way that helps us maintain enough social order so that we can sustain human dignity and flourishing.

[4] An overhead; cf. Ajay D. Kshemkalyani and Mukesh Singhal, *Distributed, Computing: Principles, Algorithms, and Systems* (Cambridge: Cambridge University Press, 2011).

AI, Including Machine Learning, Occurs by Design

Artificial intelligence only occurs by and with design. Thus AI is only produced intentionally, for a purpose, by one or more members of human society. That act of production requires design decisions concerning at a minimum the information input to and output from the system, and also where and how the computation required to transform that information will be run. These decisions entail also considerations of energy consumption and time that can be taken in producing as good a system as possible. Finally, any such system can and should be defended with levels of both cyber- and physical security appropriate to the value of the data transmitted or retained as well as the physical capacities of the system if it acts on the world.[5]

The tautology that AI is always generated by design extends to *machine learning* (ML), which is one means of developing AI wherein computation is used to discover useful regularities in data. Systems can then be built to exploit these regularities, whether to categorize them, make predictions, or select actions directly. The mere fact that part of the process of design has been automated does not mean that the system itself is not designed. The choice of an ML algorithm, the data fed into it to train it, the point at which it is considered adequately trained to be released, how that point is detected by testing, and whether that testing is ongoing if the learning continues during the system's operation—all of these things are design decisions that not only must be made but also can easily be documented. As such, any individual or organization that produces AI could always be held to account by being asked to produce documentation of these processes.

Documentation of such decisions and records of testing outcomes are easy to produce, but good practice is not always followed.[6] This is as much a matter for the law as any other sloppy or inadequate manufacturing technique.[7] The development processes deemed adequate for commercial products or even private enjoyment are determined by some combination of expertise and precedent. Whether these processes have been followed *and* documented can easily be checked either before a product is licensed, after a complaint has been made, or as a part of routine inspection.

Although actual algorithms *are* abstractions, that only means algorithms in themselves are not AI. In computer science, an algorithm is just a list of instructions to be followed, like a recipe in baking.[8] Just as a strand of DNA in itself is not life—it has no capacity to reproduce itself—so instruction sets require not only input (data) but also

[5] Note that these observations show that basic systems engineering demonstrates how uninformed the idea is of a machine converting the world into paperclips, as per Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies* (Oxford: Oxford University Press, 2014), 122–25.

[6] Michael Huttermann, *DevOps for Developers* (New York: Apress/Springer, 2012).

[7] Joshua A. Kroll et al., "Accountable Algorithms," *Univ. Penn. L. Rev.* 165 (2017): 633–706.

[8] The term algorithm is currently often misused to mean an AI system by those unclear on the distinctions between design, programs, data, and physical computing systems.

physical computation to be run. Without significant, complex physical infrastructure to execute their instructions, both DNA and AI algorithms are inert. The largest global technology corporations have almost inconceivably vast infrastructure for every aspect of storing, processing, and transmitting the information that is their business. This infrastructure includes means to generate electric power and provide secure communication as well as means to do computation.

These few leading corporations further provide these capacities also as service infrastructure to a significant percentage of the world's other ICT companies—of course, at a cost. The European Union (EU) has committed to investing substantial public resources in developing a localized equivalent of this computational infrastructure resource, as they have previously done with both commercial aviation and global positioning systems. The EU may also attempt to build a parallel data resource, though this is more controversial. There has also been some discussion of "nationalizing" significant technology infrastructure, though that idea is problematic given that the Internet is transnational. *Trans*nationalizing technology "giants" is discussed later in this chapter.

Digital technology empowers us to do all sorts of things, including obfuscating or simply deleting records or the control systems they refer to. We can make systems either harder or easier to understand using AI.[9] These are design decisions. The extent to which transparency and accountability should be required in legal products is also a design decision, though here it is legislators, courts, and regulators that design a regulatory framework. What is important to realize is that it is perfectly possible to mandate that technology be designed to comply with laws, including any that ensure traceability and accountability of the human actions involved in the design, running, and maintenance of intelligent systems. In fact, given that the limits of "machine nature" are far more plastic than those of human nature, it is more sensible to minimize the amount of change to laws and instead to maximize the extent of required compliance to and facilitation of extant laws.[10]

The Performance of Designed Artifacts Is Readily Explainable

Perhaps in the desire to evade either the laws of nations or the laws of nature, many deeply respected AI professionals have claimed that the most promising aspects of AI

[9] Kroll et al., "Accountable Algorithms."

[10] Joanna J. Bryson, Mihailis E. Diamantis, and Thomas D. Grant, "Of, For, and By the People: The Legal Lacuna of Synthetic Persons," *Artificial Intelligence and Law* 25.3 (Sept. 2017): 273–291; Margaret Boden et al., *Principles of Robotics*, The United Kingdom's Engineering and Physical Sciences Research Council (EPSRC), April 2011, https://www.epsrc.ac.uk/research/ourportfolio/themes/engineering/activities/principlesofrobotics/.

would be compromised if AI were to be regulated.[11] For example, the claim that maintaining standard rights to explanation—that is, demonstration of due process—would eliminate the utilization of many advanced machine learning techniques is based on the fact that these methods produce systems the exact workings of which are too complex to be knowable. This claim fails to take into account the present standards for accountability in corporate law. If a company is audited, that audit never extends to explaining the workings of the brain synapses or gene regulation of that company's employees. Rather, we look for audit trails—or perhaps witnesses—indicating that humans have followed appropriate procedures.

Automation exploiting artificial intelligence may reduce the number of people who can be put on a witness stand to describe their recollections of events or motivations, but it enables a standard of record keeping that would be unbearably tedious in nondigital processes. It is not the case that all AI systems are programmed to keep such records, nor that all such records are maintained indefinitely. But it *is* the case that *any* AI system can be programmed to perform such documentation, and that the programming and other development of AI can always use good systems engineering practice, including logging data on the design, development, training, testing, and operation of the systems. Further, individuals or institutions can choose how, where, and for how long to store this logged data. Again, these are design decisions for both AI systems and the institutions that create them. There are already available standards for adequate logging to generate proof of due diligence or even explanations of AI behavior. Norms of use for these or other standards can be set and enforced.[12]

What matters for human justice is that humans do the right things. We do not need to completely understand exactly how a machine-learning algorithm works any more than we need to completely understand the physics of torque to regulate bicycle riding in traffic. Our concerns about AI should be that it is used in a way that is lawful. We want to know, for example, that products comply with their claims, that individual users are not spied upon or unfairly disadvantaged, and that foreign agencies were not able to illicitly insert false information into a machine-learning dataset or a newsfeed.

All AI affords the possibility of maintaining precise accounts of when, how, by whom, and with what motivation the system deploying it has been constructed. Indeed, this is true of artifacts in general, but digital artifacts are particularly amenable to automating the process. The very tools used to build intelligent systems can also be set to capture and prompt for this kind of information. We can similarly track the construction, application, and outcomes of any validating tests. Further, even the most obscure AI system

[11] My assertion about the "deeply respected" relates to claims I've heard in high-level policy settings, but haven't been able to find in print. However, for examples of the rhetoric see Cassie Kozyrkov, "Explainable AI Won't Deliver: Here's Why," *Hackernoon* (Nov. 2018), https://hackernoon.com/explainable-ai-wont-deliver-here-s-why-6738f54216be; Cassie Kozyrkov, "The Trade-Off in Machine Learning: Accuracy vs Explain-Ability," *Medium* (Dec. 2018), https://medium.com/@erdemkalayci/the-tradeoff-in-machine-learning-accuracy-vs-explainability-fbb13914fde2.

[12] Joanna J. Bryson and Alan F. T. Winfield, "Standardizing Ethical Design for Artificial Intelligence and Autonomous Systems," *Computer* 50.5 (May 2017): 116–119.

after development can be treated entirely as a blackbox and still tested to see what variation in inputs creates variation in the outputs.[13] Even where performance is stochastic, statistics can tell us the probability of various outcomes, again a type of information to which the law is already accustomed e.g. for medical outcomes. In practice though, systems with AI are generally far less opaque than human reasoning and less complex than other problems we deal with routinely such as the workings of a government or ecosystem. There is a decades-old science of examining complex models by using simpler ones, which has been recently accelerating to serve the sectors that are already well regulated and that of course (like all sectors) increasingly use AI.[14] And of course many forms of AI, built either with or without the use of ML, do readily produce explanations themselves.[15]

To return to one of the assertions at the beginning of this section, it is also wrong to assume that AI is not already regulated. All human activity, particularly commercial activity, occurs in the context of some sort of regulatory framework.[16] The question is how to continue to optimize this framework in light of the changes in society and its capacities introduced by AI and ICT more generally.

Intelligence Increases by Exploiting Prior Computation

The fact that computation is a physical process limits how much can be done *de novo* in the instant during which intelligence must be expressed—when action must be taken to save a system from a threat or to empower it through an opportunity. For this reason, much of intelligence exploits computation already done, or rather exploits those artifacts produced that preserve the outcomes of that computation. Recognising the value and reuse of prior computation helps us understand the designs not only of culture but also of biology. Not only can organisms solely exploit opportunities they can perceive, they also tend to perceive solely what they are equipped to exploit—capacities for perception and action evolve together. Similarly, culture passes us not every tool that others have invented, but of all those inventions, the ones that produce the greatest impact relative to the costs of transmission. Costs of transmission include both time spent transmitting

[13] This process is coming to be called (as of this writing) "forensic analysis"; see, e.g., Joseph R. Barr and Joseph Cavanaugh, "Forensics: Assessing Model Goodness: A Machine Learning View," *ESCRI* 2, no. 2 (2019): 17–23.

[14] Patrick Hall, "On the Art and Science of Machine Learning Explanations," *arXiv preprint arXiv:1810.02909* (2018).

[15] Stephen Cranefield et al., "No Pizza for You: Value-based Plan Selection in BDI Agents," in *IJCAI Proceedings*, ed. Carles Sierra (Melbourne, 2017): 178–84; Jiaming Zeng, Berk Ustun, and Cynthia Rudin, "Interpretable Classification Models for Recidivism Prediction," *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 180.3 (2017): 689–722.

[16] Miles Brundage and Joanna J. Bryson, *Smart Policies for Artificial Intelligence*, in preparation, available as arXiv:1608.08196 (2017).

(reducing other opportunities) and the likelihood of inadequately faithful replication creating hazardous behaviour.[17] Culture itself evolves, and frequently those changes generate increased efficacy in those that learn them.[18]

Much of the recent immense growth of AI has been due specifically to improved capacities to "mine" using ML the prior discoveries of humanity and nature more generally.[19] Of course with such mining the good comes with the bad. We mine not only knowledge but also stereotypes—and, if we allow AI to take action, prejudice—when we mine human culture.[20] This is not a special feature of AI; as mentioned previously, this is how nature works as well.[21] Evolution can only collect and preserve the best of what is presently available (what has already been computed); even within that range the process is stochastic and will sometimes make errors. Further, examining the AI products of ML has shown that at least some of what we call "stereotypes" reflect aspects of present-day conditions, such as what proportion of job holders for a particular position have a particular gender. Thus some things we have agreed are bad (e.g. that it is sexist to expect programmers to be male) are aspects of our present culture (most programmers are male now) we have at least implicitly agreed we wish to change. Machine learning of data about present employment–or even of ordinary word use which will necessarily be impacted by present employment–cannot by itself also discover such implicit agreements and social intentions.

One theory for explaining the explosion in what we recognize as AI (that is, of AI with rich, demonstrably human-like, and previously human-specific capacities such as speech production or face recognition) is that it is less a consequence of new algorithms than of new troves of data and increased computation speeds. Where such explosions of capacities is based on the strategy of mining past solutions, we can expect that improvement to plateau. Artificial and human intelligence will come to share nearly the same boundary of extant knowledge, though that boundary will continue to expand. In fact, we can also

[17] Ivana Ćače and Joanna J. Bryson, "Agent Based Modelling of Communication Costs: Why Information Can be Free," in *Emergence and Evolution of Linguistic Communication*, ed. C. Lyon, C. L. Nehaniv, and A. Cangelosi (London: Springer, 2007), 305–322; Kenny Smith and Elizabeth Wonnacott, "Eliminating Unpredictable Variation through Iterated Learning," *Cognition* 116.3 (2010): 444–9.

[18] Alex Mesoudi, Andrew Whiten, and Kevin N. Laland, "Towards a Unified Science of Cultural Evolution," *Behavioral and Brain Sciences* 29.4 (2006): 329–47; Joanna J. Bryson, "Embodiment versus Memetics," *Mind & Soc'y* 7.1 (June 2008): 77–94; Joanna J. Bryson, "Artificial Intelligence and Pro-Social Behaviour," in *Collective Agency and Cooperation in Natural and Artificial Systems: Explanation, Implementation and Simulation*, ed. Catrin Misselhorn, vol. 122, Philosophical Studies (Berlin: Springer, 2015), 281–306; Daniel C. Dennett, *From Bacteria to Bach and Back* (London, Allen Lane, 2017).

[19] Thomas B Moeslund and Erik Granum, "A Survey of Computer Vision-based Human Motion Capture," *Computer Vision and Image Understanding* 81.3 (2001): 231–268; Sylvain Calinon et al., "Learning and Reproduction of Gestures by Imitation," *IEEE Robotics & Automation Mag.* 17.2 (2010): 44–54.

[20] Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan, "Semantics Derived Automatically from Language Corpora Contain Human-like Biases," *Sci.* 356.6334 (2017): 183–186.

[21] Molly Lewis and Gary Lupyan, "Language Use Shapes Cultural Norms: Large Scale Evidence from Gender," *Nature Human Behaviour* (accepted for publication).

expect human knowledge to be expanding faster now, given the extra computational resources we are bringing not only through digital hardware but also by our increasing access to other human minds. For humanity, ICT reduces the aforementioned overhead costs of discovering, combining, and transmitting prior computational outcomes. We all get smarter as our culture expands to embrace more—and more diverse—minds.[22] However, the fact that we can exploit our own computation to build AI, or that we can increase our own native as well as systemic intelligence by using AI, does not mean that we are replaceable with or by AI. As will be explained in the next sections, AI cannot be used to replicate humans, and this has substantial consequences for law and regulation.

AI Cannot Produce Fully Replicated Humans (All Models Are Wrong)

Computer science is often mistaken for a branch of mathematics. When this happens, many important implications of computation being a physical process are lost. For example, AI is wrongly perceived as a path toward human immortality. First, the potential of "uploading" human intelligence in any meaningful sense is highly dubious. Technologically, brains cannot be "scanned" and replicated in any other material than another brain, as their computational properties depend on trillions of temporal minutiae.[23] Creating a second, identical human to host that new brain not only is physically intractable but also would be cloning—both unethical and illegal, at least in the European Union. Second, even if we could somehow upload adequate abstractions of our own minds, we should not confuse this with actually having spawned a digital replica.[24] For example, an abstracted digital clone might be of use to manufacture canned email replies[25] or to create interactive interfaces for historical storytelling,[26] but this does not make it human.

[22] Anita Williams Woolley et al., "Evidence for a Collective Intelligence Factor in the Performance of Human Groups," *Sci.* 330.6004 (October 29, 2010): 686–688; Barton H. Hamilton, Jack A. Nickerson, and Hideo Owan, "Diversity and Productivity in Production Teams," *Advances in the Econ. Analysis of Participatory and Labor-Managed Firms* (2012): 99–138; Feng Shi et al., "The Wisdom of Polarized Crowds," *Nature Hum. Behaviour* 3 (2019): 329–336.

[23] Yoonsuck Choe, Jaerock Kwon, and Ji Ryang Chung, "Time, Consciousness, and Mind Uploading," *Int'l J. Machine Consciousness* 4.01 (2012): 257–274.

[24] As some would suggest; see Murray Shanahan, *The Technological Singularity* (Cambridge, MA: MIT Press, 2015), for a review.

[25] Mark Dredze et al., "Intelligent Email: Reply and Attachment Prediction," in *Proceedings of the 13th International Conference on Intelligent User Interfaces* (New York: ACM, 2008), 321–4.

[26] David Traum et al., "New Dimensions in Testimony: Digitally Preserving a Holocaust Survivor's Interactive Storytelling," in *Proceedings of the Eighth International Conference on Interactive Digital Storytelling* (Cham, Switzerland: Springer, 2015): 269–281.

Many have argued that the moral intuitions, motivations, even the aesthetics of an enculturated ape can in no way be meaningfully embedded in a device that shares nothing of our embodied physical ("phenomenological") experience.[27] Nothing we build from metal and silicon will ever share our phenomenology as much as a rat or cow, and few see cows or rats as viable vessels of our posterity. Yet whether such digital artifacts are viewed as adequate substitutes for a real person depends on what one values about that person. For example, for those who value their capacity to control the lives of others, many turn to the simple technology of a will to control intimate aspects of the lives of those chosen to be their heirs. It therefore seems likely that there will be those who spend millions or even billions of dollars, euros, or rubles on producing digital clones they are literally deeply invested in believing to be themselves, or at least in forcing others to treat as extensions of themselves.[28]

Even if we could somehow replicate ourselves in an artifact, the mean time for obsolescence of digital technologies and formats is far, far shorter than the average human life expectancy, which presently nears ninety years. This quick obsolescence is true not only of our physical technology but also of our fashion. Unquestionably any abstracted digital self-portrait would follow fashion in reflecting an aspect of our complex selves that will have been culturally appropriate only in a specific moment. It would not be possible from such an abstraction to fully model how our own rich individual being would have progressed through an extended lifetime, let alone through biological generations. Such complete modeling opposes the meaning of *abstraction*. An unabstracted model would again require biological cloning, but even then after many generations it would fall out of ecological fashion or appropriateness as evolution progresses.

With apologies to both Eisenhower and Box[29], all abstractions are wrong, but producing abstractions is essential. By the definition used in this chapter, all intelligence—that is, intelligent action—is an abstraction of the present context. Therefore producing an abstraction is the essence of intelligence. But that abstraction is only a snapshot of the organism; it is not the organism itself. All models are wrong, because we build them to perform actions that are not feasible using the original.

Reproducing our full organism is not required for many aspects of what is called "positive immortality."[30] Replicating our full selves is certainly not essential to writing fiction or otherwise making a lasting contribution to a culture or society, nor for having an irrevocable impact on an ecosystem. But the purpose of this chapter is to introduce AI from the perspective of maintaining social order—that is, from the perspective of

[27] Frank Pasquale, "Two Concepts of Immortality: Reframing Public Debate on Stem-Cell Research," *Yale J. L. & Hum* 14 (2002): 73–121; Bryson, "Embodiment versus Memetics"; Guy Claxton, *Intelligence in the Flesh: Why Your Mind Needs Your Body Much More Than It Thinks* (New Haven, CT: Yale University Press, 2015); Dennett, *From Bacteria to Bach and Back*.

[28] Pasquale, "Two Concepts of Immortality," questions such expenditures, or even those of in vitro fertilization, on the grounds of economic fairness.

[29] G. E. P. Box, "Robustness in the Strategy of Scientific Model Building," in *Robustness in Statistics*, ed. R. L. Launer and G. N. Wilkinson (New York: Academic Press, 1979), 201–236.

[30] Pasquale, "Two Concepts of Immortality."

law and regulation. As will be discussed in the following section, the methods for enforcing law and regulation are founded on the evolved priorities of social animals. Therefore any intelligent artifacts representing such highly abstracted versions of an individual human are not relevant to the law except perhaps as the intellectual property of their creator.

AI Itself Cannot Be Dissuaded by Law or Treaty

There is no way to ensure that an artifact could be held legally accountable.[31] Many people think the purpose of the law is to compensate, and obviously if we allow a machine to own property or at least wealth then it could in some sense compensate for its errors or misfortune. However, the law is really primarily designed to maintain social order by dissuading people from doing wrong. Law dissuades by making it clear what actions are considered wrong and then determining the costs and penalties for committing these wrong acts. This is even more true of policies and treaties, which are often constructed after long periods of negotiated agreement among peers (or at least sufficiently powerful fellow actors that more direct control is not worth its expense) about what acts would be wrong and what costs would adequately dissuade them. The Iran Nuclear Deal is an excellent example of this process.[32]

Of course all of these systems of governance can also generate revenue, which may be used by governments to some extent to right wrongs. However, none of the costs or penalties that courts can impose will matter to an AI system. We can easily write a program that says, "Don't put me in jail!" However, we cannot program the full, systemic aversion to the loss of social status and years of a finite life span, which the vast majority of humans experience as our birthright. In fact, not only humans but many social species find isolation and confinement deeply aversive—guppies can die of fright if separated from their school, and factory farming has been shown to drive pigs to exhibit symptoms of severe mental illness.[33]

We might add a bomb, camera, and timer to a robot and then program the bomb to destruct if the camera has seen no humans (or other robots) for ten minutes. Reasoning by empathy, you might think this machine is far more disuadable than a human, who can easily spend more than ten minutes alone without self destructing. But empathy is a terrible system for establishing universal ethics—it works best on those most like

[31] With no human components; Christian List and Philip Pettit, *Group Agency: The Possibility, Design, and Status of Corporate Agents* (Oxford: Oxford University Press, 2011).

[32] Kenneth Katzman and Paul K. Kerr, *Iran Nuclear Agreement*, Tech. rep. R43333, Library of Congress, Congressional Research Service, May 2016, https://crsreports.congress.gov/product/pdf/R/R43333.

[33] Françoise Wemelsfelder, "The Scientific Validity of Subjective Concepts in Models of Animal Welfare," *Applied Animal Behaviour Sci.* 53.1 (1997): 75–88.

yourself.[34] The robot's behavior could easily be utterly unaltered by this contrivance, and so it could not be said to suffer at all by the technical definitions of suffering[35], and it certainly could not be said to be dissuaded. Even if the robot could detect and reason about the consequences of its new situation, it would not feel fear, panic, or any other systemic aversion to isolation, although depending on its goals it might alter its planning to favor shorter planning horizons.

The law has been invented by—we might even say "coevolved with"—our societies in order to hold humans accountable. As an unintended consequence, only humans *can* be held accountable with our law. Even the extension of legal personality to corporations only works to the extent that real humans who have real control over those corporations suffer if the corporation does wrong. The overextension of legal personhood to a corporation designed to fail (e.g. to launder money) is known as creating a shell company. If you build an AI system and allow it to operate autonomously, it is similarly essential that you as the person who chooses to allow the system to operate autonomously will be the one who will go to jail, be fined, and so on if the AI system transgresses the law. There is simply no way to hold the AI system itself accountable or to dissuade it. Artificial intelligence being itself held accountable would be the ultimate shell company.[36]

The implicit principles that underlie our capacity to coordinate and cooperate through the law and its dissuasions have also coevolved with our complex societies. We share many of our cognitive attributes—including perception, action capacities, and, importantly, motivations—with other apes. Yet we also have specialist motivations and capacities reflecting our highly social nature.[37] No amount of intelligence in itself necessitates social competitiveness; neither does it demand acceptance by an in-group, dominance of an out-group, nor the need to achieve social status in either. These are motivations that underlie human (and other social species') cooperation and competition, that result from our evolutionary history.[38] None of this is necessary—and much of

[34] Paul Bloom, *Against Empathy: The Case for Rational Compassion* (New York: Harper Collins, 2016).

[35] Wemelsfelder, "Scientific Validity of Subjective Concepts"; Daniel C. Dennett, "Why You Can't Make a Computer That Feels Pain," *Brainstorms*, pp. 190–229 page numbers from (Cambridge, MA, MIT Press 1981, original edition: Montgomery, VT: Bradford Books, 1978), Bryson, "Artificial Intelligence and Pro-Social Behaviour"; Margaret A. Boden, "Robot Says: Whatever (The Robots Won't Take Over Because They Couldn't Care Less)," *Aeon* (August 23, 2018) (originally a lecture at the Leerhulme Centre for the Future of Intelligence), https://aeon.co/essays/the-robots-wont-take-over-because-they-couldnt-careless. Note in particular that none of the millions of currently extant robots would behave differently with these additions unless its programming was also altered (or the weight of the additions stopped it from moving.)

[36] Bryson, Diamantis, and Grant, "Of, For, and By the People."

[37] David Michael Stoddart, *The Scented Ape: The Biology and Culture of Human Odour* (Cambridge: Cambridge University Press, 1990).

[38] Stoddart, *The Scented Ape*; Ruth Mace, "The Co-evolution of Human Fertility and Wealth Inheritance Strategies," *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 353.1367 (1998): 389–397; Jillian J. Jordan et al., "Uncalculating Cooperation Is Used to Signal Trustworthiness," *Proceedings of the Nat'l Academy of Sciences* 113.31(2016): 8658–63; Simon T. Powers, Carel P. van Schaik, and Laurent Lehmann, "How Institutions Shaped the Last Major Evolutionary Transition to Large-Scale Human Societies," *Philosophical Transactions of the Royal Society B: Biological Sciences* 371.1687 (2016): 20150098.

it is even incoherent—from the perspective of an artifact. Artifacts are definitionally designed by human intent, not directly by evolution. With these intentional acts of authored human creation[39] come not only human responsibility but also an entirely different landscape of potential rewards and design constraints.[40]

AI and ICT Impact Every Human Endeavor

Given that AI can always be built to be explainable, and that only humans can be held to account, assertions that AI itself should be trustworthy, accountable, or responsible are completely misguided. If only humans can be held to account, then from a legal perspective the goal for AI transparency is to ensure that human blame can be correctly apportioned. Of course there are other sorts of transparency, such as those that support ordinary users in establishing the correct boundaries they have with their systems (defending their own interests), or for providing developers or other practitioners the ability to debug or customize an AI system.[41] Artificial intelligence can be reliable but not trustworthy—it should not require a social compact or leap of faith.[42] Consumers and governments alike should have confidence that they can determine at will who is responsible for the AI-infused systems we incorporate into our homes, our business processes, and our security.

Every task we apply our conscious minds to—and a great deal of what we do implicitly—we do using our intelligence. Artificial intelligence therefore can affect everything we are aware of doing and a great deal we have always done without intent. As mentioned earlier, even fairly trivial and ubiquitous AI has recently demonstrated that human language contains our implicit biases, and further that those biases in many cases reflect our lived realities.[43] In reusing and reframing our previous computation, AI allows us to see truths we had not previously known about ourselves, including how we transmit stereotypes,[44] but it does not automatically or magically improve us without effort. Caliskan, Bryson, and Narayanan discuss the outcome of the famous study

[39] The choice to create life through childbirth is not the same. While we may author some of child-rearing, the dispositions just discussed are shared with other primates and are not options left to parents or other conspecifics to determine.

[40] Cf. Joanna J. Bryson, "Patiency Is Not a Virtue: The Design of Intelligent Systems and Systems of Ethics," *Ethics and Info. Tech.* 20.1 (Mar. 2018): 15–26.

[41] Bryson and Winfield, "Standardizing Ethical Design."

[42] Onora O'Neill, *A Question of Trust: The BBC Reith Lectures 2002* (Cambridge: Cambridge University Press, 2002).

[43] Caliskan, Bryson, and Narayanan, "Semantics Derived Automatically from Language Corpora."

[44] Lewis and Lupyan, "Language Use Shapes Cultural Norms." Marianne Bertrand and Sendhil Mullainathan, "Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination," *Am. Econ. Rev.* 94.4 (2004): 991–1013.

showing that, given otherwise-identical resumes, individuals with stereotypically African American names were half as likely to be invited to a job interview as individuals with European American names.[45] Smart corporations are now using carefully programmed AI to avoid implicit biases at the early stages of human resources processes so they can select diverse CVs into a short list. This demonstrates that AI can—with explicit care and intention—be used to avoid perpetuating the mistakes of the past.

The idea of having "autonomous" AI systems "value-aligned" is therefore likely to be misguided. While it is certainly necessary to acknowledge and understand the extent to which implicit values and expectations must be embedded in any artifact,[46] designing for such embedding is not sufficient to create a system that is autonomously moral. Indeed, if a system cannot be made accountable, it may also not in itself be held as a moral agent. The issue should not be embedding our intended (or asserted) values in our machines, but rather ensuring that our machines allow firstly the expression of the mutable intentions of their human operators, and secondly transparency for the accountability of those intentions, in order to ensure or at least govern the operators' morality.

Only through correctly expressing our intentions should AI incidentally telegraph our values. Individual liberty, including freedom of opinion and thought, are absolutely critical not only to human well-being but also to a robust and creative society.[47] Allowing values to be enforced by the enfolding curtains of interconnected technology invites gross excesses by powerful actors against those they consider vulnerable, a threat, or just unimportant.[48] Even supposing a power that is demonstrably benign, allowing it the mechanisms for technological autocracy creates a niche that may facilitate a less-benign power—whether through a change of hands, corruption of the original power, or corruption of the systems communicating its will. Finally, who or what is a powerful actor is also altered by ICT, where clandestine networks can assemble—or be assembled—out of small numbers of anonymous individuals acting in a well-coordinated way, even across borders.[49]

Theoretical biology tells us that where there is greater communication, there is a higher probability of cooperation.[50] *Cooperation* has nearly entirely positive connotations, but

[45] Marianne Bertrand and Sendhil Mullainathan, "Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination," *American Economic Review* 94.4 (2004): 991–1013.

[46] Jeroen van den Hoven, "ICT and Value Sensitive Design," in *The Information Society: Innovation, Legitimacy, Ethics and Democracy in Honor of Professor Jacques Berleur S.J.*, ed. Philippe Goujon et al. (Boston: Springer, 2007), 67–72; Aimee van Wynsberghe, "Designing Robots for Care: Care Centered Value-Sensitive Design," *Sci. and Engineering Ethics* 19.2 (June 2013): 407–433.

[47] Julie E. Cohen, "What Privacy Is For," *Harv. L. Rev.* 126 (May 2013): 1904–1933.

[48] Brett Frischmann and Evan Selinger, *Re-engineering Humanity* (Cambridge: Cambridge University Press, 2018); Miles Brundage et al., *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation*, Tech. rep., https://maliciousaireport.com/, Future of Humanity Institute, University of Oxford, Centre for the Study of Existential Risk, University of Cambridge, Center for a New American Security, Electronic Frontier Foundation, and OpenAI, (Feb. 2018).

[49] Carole Cadwalladr, "'I Made Steve Bannon's Psychological Warfare Tool': Meet the Data War Whistleblower," *The Observer* (March 18, 2018) https://www.theguardian.com/news/2018/mar/17/data-war-whistleblower-christopher-wylie-faceook-nix-bannon-trump.

[50] Joan Roughgarden, Meeko Oishi, and Erol Akçay, "Reproductive Social Behavior: Cooperative Games to Replace Sexual Selection," *Sci.* 311.5763 (2006): 965–969.

it is in many senses almost neutral—nearly all human endeavors involve cooperation, and while these generally benefit many humans, some are destructive to many others. Further, the essence of cooperation is moving some portion of autonomy from the individual to a group.[51] The extent of autonomy an entity has is the extent to which it determines its own actions.[52] Individual and group autonomy must to some extent trade off, though there are means of organizing groups that offer more or less liberty for their constituent parts.

Many people are (falsely) preaching that ML is the new AI, and (again falsely) that the more data ML is trained on, the smarter the AI. Machine learning is actually a statistical process we use for programming some aspects of AI. Thinking that 'bigger' (more) data are necessarily better begs the question: better for what? Basic statistics teaches us that the number of data points we need to make a prediction is limited by the amount of variation in that data, providing only that the data are a true random sample of the population measured.[53] So there are natural limits for any particular task on how much data is actually needed to build the intelligence to perform it—except perhaps for surveillance. What we need for science or medicine may require only a minuscule fraction of a population. However, if we want to spot specific individuals to be controlled, dissuaded, or even promoted, then of course we want to "know all the things."[54]

The changing costs and benefits of investment at the group level that Roughgarden, Oishi, and Akçay describe has other consequences beyond privacy and liberty. Information communication technology facilitates blurring the distinction between customer and corporation; it blurs even the definition of an economic transaction. Customers now do real labor for the corporations to whom we give our custom: pricing and bagging groceries, punching data at ATMs for banks, filling in forms for airlines, and so forth.[55] The value of this labor is not directly remunerated—we assume that we receive cheaper products in return, and as such our loss of agency to these corporations might be seen as a form of bartering. "Free" services like Internet searches and email may be better understood as information bartering.[56] These transactions are not denominated with a price, which means that ICT facilitates a black or at least opaque market reducing both measured custom and therefore tax revenue. This is true for everyone who uses Internet services and interfaces, even ignoring the present controversies

[51] Bryson, "Artificial Intelligence and Pro-Social Behaviour."

[52] Harvey Armstrong and Robert Read, "Western European Micro-States and EU Autonomous Regions: The Advantages of Size and Sovereignty," *World Dev.* 23.7 (1995): 1229–1245; Maeve Cooke, "A Space of One's Own: Autonomy, Privacy, Liberty," *Philosophy & Soc. Criticism* 25.1 (1999): 22–53.

[53] Meng, Xiao-Li. "Statistical paradises and paradoxes in big data (I): Law of large populations, big data paradox, and the 2016 US presidential election." *The Annals of Applied Statistics* 12.2 (2018): 685–726.

[54] Mark Andrejevic, "Automating Surveillance," *Surveillance & Society* 17.1/2 (2019): 7–13.

[55] Bryson, "Artificial Intelligence and Pro-Social Behaviour."

[56] Joanna J. Bryson, "The Past Decade and Future of AI's Impact on Society," *Towards a New Enlightenment? A Transcendent Decade*, OpenMind BBVA (commissioned, based on a previous whitepaper for the OECD, also commissioned.), (Madrid: Taylor, 2019).

over definitions of employment raised by platforms.[57] Our failure to assign monetary value to these transactions may also explain the mystery of why AI does not seem to be increasing productivity.[58]

Artificial intelligence, then, gives us new ways to do everything we do intentionally and a great deal more. The extent to which AI makes different tasks easier and harder varies in ways that are not intuitive. This also increases and decreases the values of human skills, knowledge, social networks, personality traits, and even locations. Further, AI alters the calculations of identity and security. Fortunately, AI also gives us tools for reasoning and communicating about all these changes and for adjusting to them. But this makes group-level identity itself more fluid, complicating our ability to govern.

Who's in Charge? AI and Governance

Despite all of this fluctuation, there are certain things that are invariant to the extent of computational resources and communicative capacities. The basic nature of humans as animals of a certain size and metabolic cost, and the basic drives that determine what gives us pleasure, pain, stress, and engagement, are not altered much. How we live is and always will be enormously impacted by how our neighbors live, as we share geographically related decisions concerning investment in air, water, education, health, and security. For this reason there will always be some kind of geography-based governance. The fundamental ethical framework we have been negotiating for the last century or so of human rights is based on the responsibility of such geographically defined governments to individuals within the sphere of influence of those governments.[59] Now wise actors like the European Union have extended the notion of an individual's sovereignty over cyberassets such as personal data.[60] This makes sense for almost exactly the same reason as rights to airspace make sense. With bidirectional information access, we can influence an individual's behavior just as we could with physical force.

Recently there has been good reason to hope that we really will start mandating developers to follow best practice in software engineering.[61] If we are sensible, we will also ensure that the information systems spreading and engulfing us will also be entirely

[57] Cf. Tim O'Reilly, *WTF? What's the Future and Why It's Up to Us* (New York: Random House, 2017).

[58] Erik Brynjolfsson, Daniel Rock, and Chad Syverson, "Artificial Intelligence and the Modern Productivity Paradox: A Clash of Expectations and Statistics," *Economics of Artificial Intelligence*, Agrawal, Gans and Goldfab (eds) (Chicago: University of Chicago Press, 2017): 23–57.

[59] Sabine C. Carey, Mark Gibney, and Steven C. Poe, *The Politics of Human Rights: The Quest for Dignity* (Cambridge: Cambridge University Press, 2010).

[60] Paul Nemitz, "Constitutional Democracy and Technology in the Age of Artificial Intelligence," *Philosophical Transactions of the Royal Soc. A: Mathematical, Physical and Engineering Sciences* 376.2133 (2018): 20180089.

[61] OECD, *Recommendation of the Council on Artificial Intelligence*, OECD Legal Instruments OECD/LEGAL/0449 (includes the OECD Principles of AI) (Paris: Organisation for Economic Cooperation and Development, May 2019).

cybersecure (or else not on the Internet), with clearly documented accountability and lines of responsibility.[62] Nevertheless, even if these visions can be achieved, there are still other areas of law and governance with which we should be concerned. The last I focus on in this present chapter are the new foci of power and wealth. As just explained in the previous section, these are also parts of the "everything human" that AI and ICT are altering. Further, it is clear that achieving secure and accountable AI requires cooperation with adequate sources of power to counter those who wish to avoid the consensus of the law. Therefore wealth and power distribution, while again like cybersecurity clearly orthogonal technologically to AI, are also irrevocably intertwined with its ethical and regulated application. Problems of AI accountability and grotesquely uneven wealth distribution are unlikely to be solved independently.

In this section it should be noted that I am describing my own work in progress with colleagues,[63] but some aspects of it seem sufficiently evident to justify inclusion here. We hypothesize that when new technologies reduce the economic cost of distance, this in turn reduces the amount of easily-sustained competition in a sector. This is because locale becomes less a part of value, so higher-quality products and services can dominate ever-larger regions, up to and including in some cases the entire globe. Such a process may have sparked the gross inequality of the late nineteenth and early twentieth centuries, when rail, news and telecommunication, and oil (far easier to transport than coal or wood) were the new monopolies. Inequality spirals if capital is allowed to capture regulation, as seems recently to have happened not only with "big tech" globally but also with finance in the United Kingdom or oil in Saudi Arabia and Russia, leading to a "resource curse."[64] The early twentieth century was a period of significant havoc; in the mid-twentieth century lower inequality and political polarization cooccurred with the innovation of the welfare state, which in some countries (including the United States and United Kingdom) preceded at least World War II, though such cooperation even in these states seemed to require the motivation of the previous War and financial crash.

Governance can be almost defined by redistribution; certainly allocation of resources to solve communal problems and create public goods is governance's core characteristic.[65] Thus excessive inequality can be seen as a failure of governance.[66] Right now what we are clearly not able to govern (interestingly, on both sides of the Great Firewall of

[62] Cf. Filippo Santoni de Sio and Jeroen van den Hoven, "Meaningful Human Control over Autonomous Systems: A Philosophical Account," *Frontiers in Robotics and AI* 5 (2018): 15.

[63] Alexander J. Stewart, Nolan McCarty, and Joanna J. Bryson, "Explaining Parochialism: A Causal Account for Political Polarization in Changing Economic Environments," arXiv preprint arXiv:1807.11477 (2018).

[64] John Christensen, Nick Shaxson, and Duncan Wigan, "The Finance Curse: Britain and the World Economy," *British J. Pol. and Int'l Relations* 18.1 (2016): 255–269; Nolan M. McCarty, Keith T. Poole, and Howard Rosenthal, *Polarized America: The Dance of Ideology and Unequal Riches*, 2nd ed. (Cambridge, MA: MIT Press, 2016).

[65] Jean-Pierre Landau, "Populism and Debt: Is Europe Different from the U.S.?," Talk at the Princeton Woodrow Wilson School, and in preparation. Feb. 2016.

[66] E.g., a Gini coefficient over 0.27; Francesco Grigoli and Adrian Robles, *Inequality Overhang*, IMF Working Paper WP/17/76, International Monetary Fund, 2017. Note that too low a Gini coefficient can be problematic too.

China) are Internet companies. Perhaps similar to the market for commercial aircraft, the costs of distance are sufficiently negligible that the best products are very likely to become global monopolies unless there is a substantial government investment (e.g., the Great Firewall of China[67] or Airbus in Europe).[68] Where governance fails in a local region, such as a county, then that is also where we are likely to see political polarization and the success of populist candidates or referendum outcomes.[69]

Many problems we associate with the present moment then were not necessarily created by AI or ICT directly, but rather they were formed indirectly by facilitating increased inequality and regulatory capture. Other problems may not have been so much created as exposed by AI.[70] There are some exceptions where ICT—particularly, the capacity of digital media to be fully reproduced at a distance and to do so inexpensively—does produce qualitative change. These include changing of the meaning of ownership[71] and generating truly novel means for recognizing and disrupting human intentions, even implicit intentions not consciously known by their actors.[72] On the other hand, some things are or should be treated as invariant. As an example mentioned earlier, human rights are the painstakingly agreed foundation of international law and the obligations of a state and should be treated as core to ethical AI systems.[73]

One of the disturbing things we come to understand as we learn about algorithms is the extent to which humans are ourselves algorithmic. Law can make us more so, particularly when we constrain ourselves with it, for example with mandatory sentencing. But ordinarily, humans do have wiggle room.[74] Trust is a form of cooperation arising only in contexts of ignorance. That ignorance may be an important feature of society that ICT threatens to

[67] Roya Ensafi et al., "Analyzing the Great Firewall of China over Space and Time," *Proceedings on Privacy Enhancing Tech.* 2015.1 (2015): 61–76.

[68] Damien Neven and Paul Seabright, "European Industrial Policy: The Airbus Case," *Econ. Pol'y* 10.21 (July 1995): 313–358.

[69] Yuri M. Zhukov, "Trading Hard Hats for Combat Helmets: The Economics of Rebellion in Eastern Ukraine," Special Issue on Ukraine: Escape from Post-Soviet Legacy, *J. Comp. Econ.* 44.1 (2016): 1–15; Sascha O. Becker, Thiemo Fetzer, and Dennis Novy, "Who Voted for Brexit? A Comprehensive District-Level Analysis," *Econ. Pol'y* 32.92 (Oct. 2017): 601–650; Florian Dorn et al., "Inequality and Extremist Voting: Evidence from Germany,"Annual Conference (2018) (Freiburg, Breisgau): Digital Economy 181598, Verein für Socialpolitik / German Economic Association.

[70] Nemitz, "Constitutional Democracy and Technology in the Age of Artificial Intelligence"; Orly Mazur, "Taxing the Robots," *Pepperdine L. Rev.* 46 (2018): 277–330.

[71] Aaron Perzanowski and Jason Schultz, *The End of Ownership: Personal Property in the Digital Economy* (Cambridge, MA: MIT Press, 2016).

[72] Calo Machado and Marco Konopacki, "Computational Power: Automated Use of WhatsApp in the Brazilian Elections," *Medium* (October 26, 2018), https://feed.itsrio.org/computational-power-automated-use-of-whatsapp-in-the-elections-59f62b857033; Cadwalladr, "'I Made Steve Bannon's Psychological Warfare Tool,'"; Zhe Wu et al., "Deception Detection in Videos," *Thirty-Second AAAI Conference on Artificial Intelligence*, New Orleans, LA (2018): 16926.

[73] Philip Alston and Mary Robinson, *Human Rights and Development: Towards Mutual Reinforcement* (Oxford: Oxford University Press, 2005); David Kaye, "State Execution of the International Covenant on Civil and Political Rights," *UC Irvine L. Rev.* 3 (2013): 95–125.

[74] Cohen, "What Privacy Is For."

remove.[75] Trust allows cheating or innovating, and sometimes this may be essential. First, allowing innovation makes more tractable the level of detail about exceptions that needs to be specified. Second, of course, innovation allows us to adjust to the unexpected and to find novel, sometimes better solutions. Some—perhaps many—nations may be in danger of allowing the digital era to make innovation or free thought too difficult or individually risky, creating nationwide fragility to security threats as well as impinging on an important human right: freedom of opinion.[76] In such countries, law may bend too much toward rigidly preserving the group, and inadequately defend the individual. As I mentioned, this is not only an issue of rights but also of robustness. Individuals and variation produce alternatives–choosing among available options is a rapid way to change behavior when a crisis demonstrates change is needed.[77] Given that the digital revolution has fundamentally changed the nature of privacy for everyone, all societies will need to find a way to reintroduce and defend "wiggle room" for innovation and opinion. I believe strongly that it would be preferable if this is done not by destroying access to history, but by acknowledging and defending individual differences, including shortcomings and the necessity of learning. But psychological and political realities remain to be explored and understood, and may vary by polity.

Summary and the Robots Themselves

To reiterate my main points, when computer science is mistaken for a branch of mathematics, many important implications of computation being a physical process are lost. Further, the impact on society of the dissemination of information, power, and influence has not been adequately noted in either of those two disciplines, while in law and social sciences, awareness of technological reality and affordances has been building only slowly. Ironically, these impacts until very recently were also not much noticed in political science. Primarily, these impacts were noted only in sociology, which was unfortunately imploding at the same time AI was exploding. Similar to the myopia of computer science, psychology has primarily seen itself as studying humans as organisms. The primary ethical considerations in that field were seen as being similar to those of medical subjects, such as concerns about patient privacy. Again, some related disciplines such as media studies or marketing raised the issue, that as we better understood human behavior we might more effectively manipulate and control it, but that observation made little headway in the popular academic understanding of AI. Direct interventions

[75] O'Neill, *Question of Trust*; Paul Rauwolf and Joanna J. Bryson, "Expectations of Fairness and Trust Co-Evolve in Environments of Partial Information," *Dynamic Games and Applications* 8.4 (Dec. 2018): 891–917.

[76] Cf. Frischmann and Selinger, *Re-engineering Humanity*.

[77] Cohen, "What Privacy Is For"; Luke Stark, "The emotional Context of Information Privacy," *Info. Soc'y* 32.1 (2016): 14–27.

via neuroscience and drugs received more attention, but the potential for indirect manipulations, particularly of adults, were seemingly dismissed.

These historic errors may be a consequence of the fact that human adults are of necessity the ultimate moral agents. We are the centers of accountability in our own societies, and as such we are expected to have the capacity to take care of ourselves. The ethics of AI therefore was often reduced to its popular culture edifice as an extension of the civil rights movement.[78] Now that we have discovered—astonishingly!—that people of other ethnicities and genders are as human as "we" are, "we" are therefore obliged to consider that *anything* might be human. This position seems more a rejection of the inclusivity of civil and human rights than an appropriate extension, but it is powerfully attractive to many who seem particularly likely to be members of the recently dominant forms of gender and ethnicity, and who perhaps intuit that such an extension would again raise the power of their own clique by making the notion of rights less meaningful.

More comprehensibly, some have suggested we must extend human rights protections to anything that humans might identify with in order to protect our own self-concept, even if our identification with these objects is implicit or mistaken.[79] This follows from Kant's observation that those who treat animals reminiscent of humans badly are also more likely to treat humans badly. Extending this principle to AI though is most likely also a mistake, and an avoidable one. Remember that AI is definitionally an artifact and therefore designed. It almost certainly makes more sense where tractable to change AI than to radically change the law. Rather than Kant motivating us to treat AI that appears human as if it were human, we can use Kant to motivate not building AI to appear human in the first place. This has been the approach of first the United Kingdom[80] and now very recently the OECD[81] whose AI ethics principles recommend that AI should never deceptively appear to be human. This may seem like a heavy restrictiction at present, but as society becomes more familiar with AI—and, through that process, better understands what it is about being human that requires and deserves protection—we should be able to broaden the scope of how humanlike devices can be while still not having that likeness deceive.[82]

There are recent calls to ground AI governance not on "ethics" (which is viewed as ill-defined) but on international human rights law. Of course, this may be a false dichotomy; procedures from classical ethics theories may still be of use in determining

[78] Tony J. Prescott, "Robots Are Not Just Tools," *Connection Sci.* 29.2 (2017): 142–149; David J. Gunkel, "The Other Question: Can and Should Robots Have Rights?," *Ethics and Info. Tech.* 20.2 (2018): 87–99; Daniel Estrada, "Value Alignment, Fair Play, and the Rights of Service Robots," *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES 20'18, New York, NY, ACM (2018), 102–107.

[79] Joel Parthemore and Blay Whitby, "What Makes Any Agent a Moral Agent? Reflections on Machine Consciousness and Moral Agency," *Int'l J. Machine Consciousness* 5.02 (2013): 105–129; David J. Gunkel, *Robot Rights* (Cambridge, MA: MIT Press, 2018).

[80] Boden et al., *Principles of Robotics*.

[81] OECD, *Recommendation of the Council on Artificial Intelligence*.

[82] Joanna J. Bryson, "The Meaning of the EPSRC Principles of Robotics," *Connection Sci.* 29.2 (2017): 130–136.

ambiguities and trade-offs of law's application.[83] We can certainly expect ongoing consideration of localized variation, which the term *ethics* perhaps better communicates than *rights*. Ethics has always been about identity communicated in codes of conduct, which confound fundamental principles that we may be able to codify as rights with other things that are essentially identity markers. But identity too can be essential to security through constructing a defendable community.[84] Identity obviously (definitionally) defines a group, and groups are often the best means humans have for achieving security and therefore viability. Not only is breaking into different groups sometimes more efficient for governance or other resource constraints, but also some groups will have different fundamental security trade-offs based on their geological and ecological situation or just simply their relations with neighbors. Identity also often rests on shared historical narratives, which afford different organizational strategies. These of course may be secondary to more essential geo-ecological concerns, as is illustrated by the apparent ease with which new ethnicities are invented.[85] All of these of course also make a contribution to security, and get wrapped up in localised ethical systems.

In conclusion, any artifact that transforms perception to more relevant information, including action, is AI—and note that AI is an adjective, not a noun, unless it is referring to the academic discipline. There is no question that AI and digital technologies more generally are introducing enormous transformations to society. Nevertheless, these impacts should be governable by less transformative legislative change. The vast majority of AI—particularly where it has social impact—is and will remain a consequence of corporate commercial processes, and as such subject to existing regulations and regulating strategies. We may need more regulatory bodies with expertise in examining the accounts of software development, but it is critical to remember that what we are holding accountable is not the machines themselves but the people who build, own, or operate them—including any who alter their operation through assault on their cybersecurity. What we need to govern is the human application of technology, and what we need to oversee are human processes of development, testing, operation, and monitoring.

Artificial intelligence also offers us an opportunity to discover more about how we ourselves and our societies work. By allowing us to construct artifacts that mimic aspects of nature but provide new affordances for modularity and decoupling, we allow ourselves novel means of self-examination, including examination of our most crucial capacities such as morality and political behavior. This is an exciting time for scientific and artistic exploration as well as for commerce and law. But better knowledge also

[83] Cansu Canca, "Human Rights and AI Ethics: Why Ethics Cannot Be Replaced by the UDHR," *United Nations Univ.: AI & Global Governance Articles & Insights* (July 2019), https://cpr.unu.edu/ai-global-governance-human-rights-and-ai-ethics-why-ethics-cannot-be-replaced-by-the-udhr.html.

[84] Bill McSweeney, *Security, Identity and Interests: A Sociology of International Relations* Cambridge University Press (1999); Simon T. Powers, "The Institutional Approach for Modeling the Evolution of Human Societies," *Artif. Life* 24.1 (2018): 10–28.

[85] Erin K. Jenne, Stephen M. Saideman, and Will Lowe, "Separatism as a Bargaining Posture: The Role of Leverage in Minority Radicalization," *J. Peace Research* 44.5 (2007): 539–558.

offers an opportunity for better control. The role of the law for crafting both individual and societal protections has never been more crucial.

Acknowledgments

A small proportion of the material in this review was derived from a document previously delivered to the OECD (Karine Perset) in May 2017 under the title "Current and Potential Impacts of Artificial Intelligence and Autonomous Systems on Society," which contributed to the OECD AI policy efforts and documents of 2018–2019, and also was reused (with permission) and expanded for Bryson (2019 BBVA). More debt is probably owed to Frank Pasquale for extremely useful feedback and suggestions on a first draft. Thanks also to Will Lowe, Patrick Slavenburg, and Jean-Paul Skeete. I was supported in part by an AXA Research Fellowship in AI Ethics while writing this chapter.

References

Boden, Margaret et al. *Principles of Robotics*. The United Kingdom's Engineering and Physical Sciences Research Council (EPSRC). Apr. 2011. https://www.epsrc.ac.uk/research/ourportfolio/themes/engineering/activities/principlesofrobotics/.

Brundage, Miles et al. *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation*. Tech. rep. https://maliciousaireport.com/. Future of Humanity Institute, University of Oxford, Centre for the Study of Existential Risk, University of Cambridge, Center for a New American Security, Electronic Frontier Foundation, and OpenAI, Feb. 2018.

Bryson, Joanna J. "The Past Decade and Future of AI's Impact on Society." In *Towards a New Enlightenment? A Transcendent Decade*, OpenMind BBVA (commissioned, based on a white paper also commissioned, that by the OECD). Madrid: Taylor, Mar. 2019.

Bryson, Joanna J., Mihailis E. Diamantis, and Thomas D. Grant. "Of, For, and by the People: The Legal Lacuna of Synthetic Persons." *Artificial Intelligence and Law* 25.3 (Sept. 2017): 273–91.

Cadwalladr, Carole. "'I Made Steve Bannon's Psychological Warfare Tool': Meet the Data War Whistleblower." *The Observer* (March 18, 2018).

Claxton, Guy. *Intelligence in the Flesh: Why Your Mind Needs Your Body Much More Than It Thinks*. New Haven, CT: Yale University Press, 2015.

Cohen, Julie E. "What Privacy Is For." In: *Harv. L. Rev.* 126 (May 2013): 1904–33.

Dennett, Daniel C. "Why You Can't Make a Computer That Feels Pain." *Brainstorms*. Reprint. Montgomery, VT: Bradford Books, 1978, 190–229.

Gunkel, David J. *Robot Rights*. Cambridge, MA: MIT Press, 2018.

Hüttermann, Michael. *DevOps for Developers*. New York: Apress/Springer, 2012.

Kroll, Joshua A., et al. "Accountable Algorithms." *Univ. Penn. L. Rev.* 165 (2017): 633–706.

List, Christian, and Philip Pettit. *Group Agency: The Possibility, Design, and Status of Corporate Agents*. Oxford: Oxford University Press, 2011.

Nemitz, Paul. "Constitutional Democracy and Technology in the Age of Artificial Intelligence." *Philosophical Transactions of the Royal Soc. A: Mathematical, Physical and Engineering Sciences* 376.2133 (2018): 20180089.

OECD. *Recommendation of the Council on Artificial Intelligence*. OECD Legal Instruments OECD/LEGAL/0449 (includes the OECD Principles of AI). Paris: Organisation for Economic Cooperation and Development, May 2019.

O'Neill, Onora. *A Question of Trust: The BBC Reith Lectures 2002*. Cambridge: Cambridge University Press, 2002.

O'Reilly, Tim. *WTF? What's the Future and Why It's Up to Us*. New York: Random House, 2017.

Santoni deSio, Filippo, and Jeroen van den Hoven. "Meaningful Human Control over Autonomous Systems: A Philosophical Account." *Frontiers in Robotics and AI* 5(2018): 15.

Shanahan, Murray. *The Technological Singularity*. Cambridge, MA: MIT Press, 2015.

Sipser, Michael. *Introduction to the Theory of Computation*. 2nd ed. Boston: PWS, Thompson, 2005.