EDITED BY

MARKUS D.
DUBBER

FRANK
PASQUALE

SUNIT
DAS

The Oxford Handbook *of* ETHICS OF AI

THE OXFORD HANDBOOK OF

ETHICS OF AI

Edited by

MARKUS D. DUBBER, FRANK PASQUALE,

and

SUNIT DAS

OXFORD
UNIVERSITY PRESS

OXFORD
UNIVERSITY PRESS

Oxford University Press is a department of the University of Oxford. It furthers the University's objective of excellence in research, scholarship, and education by publishing worldwide. Oxford is a registered trade mark of Oxford University Press in the UK and certain other countries.

Published in the United States of America by Oxford University Press
198 Madison Avenue, New York, NY 10016, United States of America.

Library of Congress Control Number: 2020938746

ISBN 978-0-19-006739-7

1 3 5 7 9 8 6 4 2

Printed by Sheridan Books, Inc., United States of America.

Contents

PART I. INTRODUCTION AND OVERVIEW

PART II. FRAMEWORKS AND MODES

PART III. CONCEPTS AND ISSUES

PART IV. PERSPECTIVES AND APPROACHES

PART V. CASES AND APPLICATIONS

Editors' Preface

The idea for this handbook arose in late 2017, with the working title *Handbook of Ethics of AI in Context*. By the time solicitations went out to potential contributors in the summer of 2018, its title had been streamlined to *Handbook of Ethics of AI*. Its essentially contextual approach, however, remained unchanged: it is a broadly conceived and framed interdisciplinary and international collection, designed to capture and shape much-needed reflection on normative frameworks for the production, application, and use of artificial intelligence in diverse spheres of individual, commercial, social, and public life.

The approach to the ethics of AI that runs through this handbook is contextual in four senses:

- it locates ethical analysis of artificial intelligence in the context of other modes of normative analysis, including legal, regulatory, philosophical, and policy approaches,
- it interrogates artificial intelligence within the context of related modes of technological innovation, including machine learning, Big Data, and robotics,
- it is interdisciplinary from the ground up, broadening the conversation about the ethics of artificial intelligence beyond computer science and related fields to include other fields of scholarly endeavor, including the social sciences, humanities, and the professions (law, medicine, engineering, etc.), and
- it invites critical analysis of all aspects of—and participants in—the wide and continuously expanding artificial intelligence complex, from production to commercialization to consumption, from technical experts to venture capitalists to self-regulating professionals to government officials to the general public.

Ideally, handbooks combine stock-taking and genre-defining. Devoted to a field of inquiry as new and quickly evolving as ethics of AI, this handbook falls closer to the forward-facing than to the literature-reviewing end of the spectrum. Mapping the existing discourse is important, also as the beginning of a crucial attempt to place current developments in historical context. At the same time, we recognized the need to leave room for flexibility as the contributors to this volume broke new ground, pursuing fresh approaches and taking on novel subjects. In the same spirit, this handbook operates with an inclusive and flexible conception of "artificial intelligence" that ranges from exploring normative constraints on specific applications of machine learning algorithms to reflecting on the (potential) status of AI as a form of consciousness with attendant rights and duties and, more generally still, to investigating the basic conceptual terms and frameworks necessary to understand tasks requiring intelligence, whether "human" or "AI."

Each chapter in this handbook aims to provide an original, critical, and accessible account of the current state of debate in its domain that will help to shape scholarly research and public discourse. We have welcomed forward-looking and ideas-driven contributions, to serve as catalysts for guiding the debate on the ethics of AI in the months and years to come. The chapters are intended to function, individually and collectively, as lively, freestanding essays targeted at an international and interdisciplinary audience of scholars and interested laypersons. Each chapter also provides, at the end, a bibliography of about ten titles for readers who would like to read more deeply into the topic.

The handbook's inclusive and flexible approach to its subject matter is reflected in its roster of contributors, which includes authors from several countries and continents, ranging from emergent to established authorities and representing a wide variety of methodological approaches, areas of expertise, and research agendas. The handbook's content is similarly ambitious and diverse in scope and substance, covering a broad range of topics and perspectives. The handbook consists of five parts: I. Introduction and Overview, II. Frameworks and Modes, III. Concepts and Issues, IV. Perspectives and Approaches, and V. Cases and Applications.

Part I provides a general introduction to the subject (and field) of "artificial intelligence" within the context of research and discourse in related fields of technological innovation, laying an accessible yet nuanced foundation for the exploration of various normative frameworks for the critical analysis of AI. It also locates the "ethics" of artificial intelligence in relation to cognate fields of ethical inquiry (e.g., data ethics, information ethics, robot ethics, internet ethics), considering ways of conceptualizing it and its challenges (e.g., as a sui generis inquiry, as a form of applied ethics, or as traditional ethics in AI terms), distinguishing aspects within it (to the extent a taxonomy of this sort proves illuminating), and capturing some key substantive and formal features of the discourse.

Part II places the subject of this handbook, the ethics of AI, within the context of alternative frameworks for normative assessment and governance, including various institutional and procedural modes of implementation and dissemination. Questions raised in this part include: "What distinguishes the ethics of AI from other normative frameworks and techniques, e.g., law, policy, regulation, governance?"; "How can ethics ground and inform legal constraints on (and regulatory guidance for) AI?"; "How does an ethics of AI navigate the possible tension between private commercial norms, on the one hand, and public norms, on the other?"; "How should ethical norms be generated and formulated, disseminated and implemented, and by whom?"; and "What is the role of the (self-)regulation of professional ethics, insofar as this enterprise is regarded as defining and enforcing a notion of good, sound, or 'professional' judgment?"

Part III tackles central concepts and issues that may serve as points of departure for reflecting on the ethical dimensions and challenges of artificial intelligence in general, cutting across technologies and applications, and in many cases across disciplines as well, ranging from the sources and types of bias in the production and application of AI research, to concerns about privacy in the collection and use of data, the potential effect of AI-driven "disruption" on labor markets and the future of work and on socioeconomic life more broadly, the distinction between "prediction" and "judgment," and the ethical status of AI-driven machines and its possible implications for human-machine interaction.

While a wide spectrum of disciplinary, national, and supranational perspectives is reflected throughout the handbook, Part IV homes in on a selection of methodological approaches and domestic or regional contexts. Early chapters in this part capture the distinctive texture and salience of actual (or potential) discourse around ethics of AI in a range of disciplinary contexts, in an effort to illustrate—and to expand—the disciplinary scope of the scholarly and public debate about ethics of AI. The remaining chapters highlight the variety of discourses around ethics of AI in selected national and regional contexts, again to broaden and to diversify the dialogue about the normative dimensions of artificial intelligence as a global phenomenon, this time geographically and culturally.

Part V concludes the handbook by sharpening its focus to selected applications of artificial intelligence, without, however, treating them as sui generis, but instead in a way that fits into the handbook's overall ambition: to expand the conversation about the ethics of artificial intelligence from the specific to the general, from the superficial to the fundamental, and from the parochial to the contextual. Contributors here reflect on the ethical aspects of the design, dissemination, and use of AI-driven devices and tools today and in the future, along a broad spectrum of applications, in health care, law, immigration, education, transportation, the military, the workplace, smart cities, and beyond.

We are deeply grateful to the international and interdisciplinary group of scholars who signed on to this large-scale long-term project and somehow made the time to see it through to completion, among the flurry of activities and opportunities that mark the start of a new and momentous endeavor like the scholarly and public scrutiny of the ethics of artificial intelligence.

Markus D. Dubber, Frank Pasquale, and Sunit Das

August 2019

List of Contributors

Ifeoma Ajunwa, Associate Professor, Labor Relations, Law, and History Department, Cornell University ILR School; Associate Faculty Member, Cornell Law School; Faculty Associate, Berkman Klein Center for Internet & Society at Harvard University

Chinmayi Arun, Resident Fellow, Information Society Project at Yale Law School; Affiliate of the Berkman Klein Center for Internet & Society at Harvard University; Assistant Professor of Law, National Law University, Delhi

Benjamin R. Baer, Department of Statistics and Data Science, Cornell University

Chelsea Barabas, Massachusetts Institute of Technology

John Basl, Department of Philosophy and Religion, Northeastern University

Alessandro Blasimme, Health Ethics and Policy Lab, Department of Health Sciences and Technology, Swiss Federal Institute of Technology—ETH Zurich

Paula Boddington, New College of the Humanities, London

Joseph Bowen, Department of Philosophy, University of St Andrews; Department of Philosophy, University of Stirling

Kiel Brennan-Marquez, University of Connecticut School of Law

Joanna J. Bryson, Professor of Ethics and Technology, Hertie School for Governance

Ron Chrisley, Visiting Scholar, Institute for Human-Centered Artificial Intelligence; Visiting Professor, Symbolic Systems Program, Stanford University

John Danaher, Senior Lecturer, School of Law, National University of Ireland, Galway

Nicholas Diakopoulos, Northwestern University School of Communication

Virginia Dignum, Department of Computing Science, Umeå University

Judith Donath, Berkman Klein Center for Internet & Society at Harvard University

Elizabeth Edenberg, Assistant Professor, Baruch College, The City University of New York

Danit Gal, Technology Advisor to the UN Secretary General High-Level Panel on Digital Cooperation

Jai Galliott, Director, Values in Defence & Security Technology Group, University of New South Wales at Australian Defence Force Academy; Non-Resident Fellow, Modern War Institute at the United States Military Academy, West Point; Visiting Fellow, Centre for Technology and Global Affairs, University of Oxford

Jean-Gabriel Ganascia, Professor of Computer Science, Sorbonne University; LIP6 Laboratory—ACASA Group Leader

Urs Gasser, Executive Director, Berkman Klein Center for Internet & Society at Harvard University; Professor of Practice, Harvard Law School

Timnit Gebru, Senior Research Scientist, Google; Co-founder and President, Black in AI

Daniel E. Gilbert, Department of Statistics and Data Science, Cornell University

Ellen P. Goodman, Professor, Rutgers Law School; Co-director, Rutgers Institute for Information Policy & Law

David J. Gunkel, Department of Communication, Northern Illinois University

Andrew Howes, Professor of Computer Science, University of Birmingham

Meg Leta Jones, Communication, Culture & Technology Department, Georgetown University

Mark Kingwell, Professor of Philosophy, University of Toronto

Anton Korinek, Department of Economics and Darden School of Business, University of Virginia

Joshua A. Kroll, Department of Computer Sciences, Naval Postgraduate School

Benjamin Kuipers, Professor of Computer Science and Engineering, University of Michigan

Matthew Le Bui, Annenberg School for Communication and Journalism, University of Southern California

Karen Levy, Department of Information Science, Cornell University

Shannon Mattern, Department of Anthropology, The New School for Social Research, New York

Jason Millar, Assistant Professor, School of Electrical Engineering and Computer Science, and Canada Research Chair in the Ethical Engineering of Robotics and AI, University of Ottawa

Petra Molnar, University of Toronto Faculty of Law

Pegah Moradi, Department of Information Science, Cornell University

Deirdre K. Mulligan, Associate Professor, School of Information; Faculty Director, Berkeley Center for Law and Technology, University of California, Berkeley

Helen Nissenbaum, Professor, Information Science, Cornell Tech

Safiya Umoja Noble, Departments of Information Studies and African American Studies, University of California, Los Angeles

Ganna Pogrebna, Professor of Behavioural Economics and Data Science, University of Birmingham; Lead for Behavioural Data Science, Alan Turing Institute

Thomas M. Powers, Department of Philosophy, University of Delaware

Andrea Renda, Senior Research Fellow and Head of Global Governance, Regulation, Innovation and the Digital Economy, CEPS, Brussels; Professor of Digital Innovation, College of Europe, Bruges; Member of the EU High Level Expert Group on Artificial Intelligence

Kathleen Richardson, Professor of Ethics and Culture of Robots and AI, School of Computer Science and Informatics, De Montfort University

Nagla Rizk, Professor of Economics, The American University in Cairo

Rachel Schlund, Department of Organizational Behavior, Cornell University

Carolyn Schmitt, Berkman Klein Center for Internet & Society at Harvard University

Susan Schneider, Associate Professor of Philosophy and Director of the AI, Mind, and Society Group, University of Connecticut

Jason Scholz, Chief Executive Officer, Trusted Autonomous Systems Defence Cooperative Research Centre, Australia

Avery Slater, Assistant Professor of English, University of Toronto

Tom Slee, SAP Canada

Bryant Walker Smith, Associate Professor of Law and (by courtesy) Engineering at the University of South Carolina; Affiliate Scholar at the Center for Internet and Society at Stanford Law School; Co-director of the Program on Law and Mobility at the University of Michigan Law School; http://newlypossible.org

Norman W. Spaulding, Sweitzer Professor of Law, Stanford University

Harry Surden, Associate Professor of Law, University of Colorado

Cody Turner, AI, Mind, and Society Group, University of Connecticut

Effy Vayena, Health Ethics and Policy Lab, Department of Health Sciences and Technology, Swiss Federal Institute of Technology—ETH Zurich

Martin T. Wells, Department of Statistics and Data Science, Cornell University

Michael Wheeler, Professor of Philosophy, University of Stirling

Karen Yeung, Interdisciplinary Professorial Fellow in Law, Ethics, and Informatics at Birmingham Law School and the School of Computer Science at the University of Birmingham

Elana Zeide, PULSE Fellow in Artificial Intelligence, Law, and Policy at the University of California, Los Angeles School of Law

PART I

INTRODUCTION AND OVERVIEW

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 underinformed 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 Čače 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] Caio 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.

CHAPTER 2

THE ETHICS OF THE ETHICS OF AI

THOMAS M. POWERS AND
JEAN-GABRIEL GANASCIA

Introduction

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

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

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

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

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

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

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

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

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

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

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

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

Conceptual Ambiguities

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

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

Agent

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

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

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

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

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

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

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

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

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

Autonomy

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

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

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

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

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

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

Intelligence

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

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

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

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

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

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

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

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

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

Risk: Overestimation and Underestimation

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

Overestimations and Existential Threats from AI

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Underestimation of AI Risks

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Implementing Ethics

Making Machines Moral

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

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

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

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

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

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

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

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

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

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

Modeling Ethical Reasoning

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

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

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

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

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

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

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

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

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

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

Learning Values

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

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

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

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

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

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

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

Intrinsic Limitations

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

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

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

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

Epistemic Issues with Ethical Implications: Predictive Science

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

Artificial intelligence enters the concern about epistemology in ethics in virtue of the fact that AI is an increasingly large "supplier" of scientific information and results—especially in those disciplines identified as practicing Big Data science—and as AI continues to grow in importance for science, our epistemic dependence on AI will only increase. This will be true of descriptions of the natural world, but also of predictions, since they come from data-intensive mathematical models. So another important challenge for the ethics of ethics of AI is how AI is increasingly used to establish scientific facts, and whether those facts can be readily explained either to the lay public or in some cases even to expert scientists themselves. Here we focus on ways in which AI might create a future body of scientific results that will fall short of adding to our scientific understanding. The problem is a peculiar feature of AI in that there can be considerable generated knowledge (in terms of correlations of data and phenomena), but no commensurate increase in genuine human scientific understanding.

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

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

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

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

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

The Crisis of Causal Knowledge

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

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

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

[57] Judea Pearl, "Causal Inference in Statistics: An Overview," *Statistics Surveys* 3 (2009): 96–146.

[58] Nancy Cartwright, "A Philosopher's View of the Long Road from RCTs to Effectiveness," *The Lancet* 377:9775 (2011): 1400–1401.

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

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

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

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

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

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

[59] Pearl, "Causal Inference in Statistics," 99.

[60] David Weinberger, *Too Big to Know* (New York: Basic Books, 2011).

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

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

How an Epistemic Crisis Could Become an Ethical Crisis

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

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

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

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

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

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

Oppositional versus Systemic Approaches

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

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

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

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

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

Conclusion

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

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

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

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

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

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

Bibliography

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

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

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

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

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

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

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

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

CHAPTER 3

ETHICAL ISSUES IN OUR RELATIONSHIP WITH ARTIFICIAL ENTITIES

JUDITH DONATH

Introduction

This chapter is about the ethics of our relationships with artificial entities—bots, robots, and other computational systems created to interact with us as if they were sentient and autonomous individuals. They may be embodied as robots or exist only in software; some are clearly artificial while others are indistinguishable, at least under certain conditions, from human beings. When are such interactions helpful or harmful? How do our relationships with computational entities change our relationships with other human beings? When does it matter if we interact with a machine or a human, and why?

Sentience—the ability to have emotions, to feel pain and want to avoid it—is a core concept here. We have ethical responsibilities to sentient beings that we do not have to nonsentient objects: it is cruel to kick a dog, but not a rock. While actually sentient artificial entities might someday exist, they are as yet only a theoretical possibility. All currently existing artificial entities are nonsentient, but—unlike a rock—their interactions and designs evoke the impression of conscious entities with personalities and emotions.

Simulated sentience is the primary focus of this chapter, highlighting our relationship with entities that appear to be sentient but are not. Some are quite simple; our tendency toward anthropomorphism can make the output of even primitive programs appear to us as the behavior of a cognizant mind. Others are impenetrably complex, with sophisticated imitations of conscious and intelligent behavior that are nearly impossible to distinguish from the actions of an actually conscious being.

Some of the ethical issues we will examine involve our personal relationships with artificial entities. People seek companionship from artificial assistants, hold funeral services for broken robot dogs, and confide in simulated therapists. The relationships that some warn are a threat to humaneness, if not to humanity, are proving to be quite popular. Under what circumstances are they helpful or harmful? How do such human/machine interactions affect our relationships with other people? How does the machine performance of emotion differ from human impression management or from the inauthentic expression required by, for example, the service industry? When and why does it matter that the other does not actually think? The key issues here concern empathy and the function that caring what others think plays in society.

We will also address ethical issues in the design and deployment of artificial entities. In their mimicry of sentient beings, artificial entities are inherently deceptive: even one that types "I am a bot" implies, with its first-person pronoun, a self-conscious being. And many artificial entities are designed to be as persuasive as possible, eliciting affection and trust with features such as big childlike eyes and imitative gestures. Some are made with beneficial goals—to serve the user as teacher, wellness coach, etc.—but these same persuasive techniques can manipulate us for harmful and exploitive ends. What are the ethical responsibilities of researchers and designers?

While some artificial entities attempt to pass as human, many are clearly robots or software agents; the illusion they project is of a sentient but also distinctly artificial being. Yet the popular vision of truly sentient machine beings is generally foreboding—they are often portrayed as a potent, if not the final, enemy of humanity. Why do we see this future so darkly? While understanding the ethical issues surrounding our relationship with artificial entities is important in itself as social robots and software agents become increasingly present in our everyday lives, these queries also shed revealing light on our relationships with each other and with other living things.

Scope and Definitions

We will start with some definitions. Much discussion about today's nonsentient social robots and programs uses language that implies they have feelings and intentions, blurring the important distinction between "X is a robot that feels" and "X is a robot designed to appear as if it feels." Having a clear understanding of what is meant by intelligence, sentience, and consciousness and using them precisely is important for many ethical considerations.

Intelligence is often described as the ability to learn and apply knowledge or to solve complex problems.[1] It is an observable property defined by behavior—finding clever solutions, acting resourcefully. Thought of this way, we see a migrating bird, an insect-hunting bat, and a theorem-proving human as problem solvers each of whom require considerable, albeit very different forms of, intelligence. Thought of this way, we can easily refer to a machine as intelligent if it solves difficult problems. In this usage, the internal state that produces the intelligent behavior does not matter.

[1] Max Tegmark, "Let's Aspire to More Than Making Ourselves Obsolete, "*Possible Minds: Twenty-Five Ways of Looking at AI*, ed. John Brockman (New York: Penguin, 2019), 76–87.

Yet intelligence is not a precisely defined term.[2] It is sometimes conceptualized as an inner quality, as when we say the migrating bird is not really intelligent, but is just acting on instinct. Computer scientists joke that use of the term "artificial intelligence" also reflects this enigmatic property: computer programs that solve complex problems using methods we do not understand are "artificial intelligence"; when we do understand them they are "algorithms."

Sentience is the ability to experience sensations and emotions: to feel pain and pleasure, and to want less of the former and more of the latter. A nonsentient creature may move away from certain things and toward others, and even have a suite of behaviors that aid its survival and reproduction, but it is not motivated to do anything: it simply exists. With sentience comes motivation: a creature that experiences certain sensory inputs as painful will want to avoid those; it will want to repeat pleasant ones. Sentience is now believed to be the foundation of learning, which gives sentient creatures much greater flexibility in their relationship with the world.[3]

Sentience is central to ethics because we have responsibilities toward sentient beings that we do not have toward, say, a rock.[4] Most people would agree that we should not inflict needless pain on something capable of experiencing distress. However, which beings are included in that category and what to do when that responsibility conflicts with other needs and desires are highly contested questions.

The term *conscious* refers to sentient beings that are self-aware—that have a sense of purpose and of themselves as individuals in the world. The term can be fuzzy: there is no clear behavioral marker of consciousness nor even an agreed-upon description of the internal experience. Historically, the rationalist, Enlightenment view was that consciousness was the affectless mental acquisition and manipulation of a symbolic representation of the world. Some believed that it required language and thus humans were the only conscious animal. Today, consciousness is increasingly understood to have evolved through social interaction, beginning with the bonding of parent and offspring; it is built on the emotional scaffolding of sentience.[5] And ethological and neuroscientific studies affirm that humans are far from being the only conscious animal: many mammals, birds, even cephalopods are aware of themselves and others and move through life with intentions.[6]

[2] Shane Legg and Marcus Hutter, "Universal Intelligence: A Definition of Machine Intelligence," *Minds and Machines* 17, no. 4 (2007): 391–444.

[3] Zohar Z. Bronfman, Simona Ginsburg, and Eva Jablonka, "The Transition to Minimal Consciousness through the Evolution of Associative Learning," *Frontiers in Psychology* 7 (2016): 1954.

[4] Donald M. Broom, *Sentience and Animal Welfare* (Wallingford, UK: CABI, 2014); Peter Singer, *Practical Ethics* (Cambridge, UK: Cambridge University Press, 2011).

[5] Tania Singer et al., "Empathy for Pain Involves the Affective but not Sensory Components of Pain," *Science* 303, no. 5661 (2004): 1157–1162.

[6] Evan Thompson, "Empathy and Consciousness." *Journal of Consciousness Studies* 8, nos. 5–6 (2001): 1–32; Jaak Panksepp, "Affective Consciousness: Core Emotional Feelings in Animals and Humans." *Consciousness and Cognition* 14, no. 1 (2005): 30–80; Peter Godfrey-Smith, *Other Minds: The Octopus and the Evolution of Intelligent Life* (London: William Collins, 2016).

These differing views of what consciousness is have important repercussions for ethics and AI. In the classical view—which remains influential in some AI research as well as popular belief—consciousness is closely entwined with intelligence, the acquisition of knowledge, and problem solving. This contrasts sharply with the biological view, supported by current research, that consciousness is fundamentally social and emotional, having evolved from simple sentience as creatures began to bond and care for each other.

Consciousness is important in ethics because the basis of morality is here, in the evolution of traits such as attachment, empathy, and the desire for justice and social order. To care about how one is perceived by others and about one's effect on them—concerns available to the conscious mind—is arguably the very foundation of ethics.

Both sentience and consciousness are inherently private experiences. We cannot directly experience what it is like to be another being—human, animal, or robot. Our assessment of what it is like to be another, including what, if anything, they feel, is based on external and perceivable appearance and behavior. I assume other people are conscious because I know that I am conscious and we are biologically and behaviorally similar; it is, however, an assumption and not direct knowledge.

As we look at other species (or artificial entities), we make inferences about what it is like to be them—what their internal experience is—by analogy. The more something resembles ourselves, the more we assume his, her, or its experience to be similar to our own. This rule of thumb has led us to vastly underestimate the cognitive ability and sensate experience of many nonhuman animals and, as we shall see, to overestimate the capabilities of bots and other nonsentient human inventions.

Precursors: Turing and Weizenbaum

Our inability to directly observe the experience of being another is the problem at the core of Alan Turning's 1950 paper, "Computing Machinery and Intelligence," that marks the beginning of the field of artificial intelligence.[7] Turing introduced the paper by saying, "I propose to consider the question, 'Can machines think?'" and then immediately rejected the question on the basis that the words "machine" and "think" were too vague and limited by everyday experience.

Instead, he proposed a test, the Imitation Game, now popularly known as the Turing Test, which he argued was a "more accurate form of the question." In this test a human judge chats (via text) with two hidden contestants. Both claim to be human, though only one is—the other is a machine. The judge is tasked with determining which one is telling the truth. A machine that can consistently pass as human, Turing argued, should be considered intelligent.

[7] Alan Turing, "Computing Machinery and Intelligence." *Mind* 49 (1950): 433–60.

It is a peculiar article and a hugely influential one.[8] It anointed deceptively passing as human as the key goal—or even as the definition of—artificial intelligence. And it deftly limited the domain in which this goal needed to be achieved to text-only communication.

Turing famously predicted that in fifty years computers would have reached the point that they would be consistently able to fool a human judge.[9] But he also made a second prediction: that by the time computers could pass as human, our use of language would have changed significantly. He said, "The original question, 'Can machines think' I believe to be too meaningless to deserve discussion. Nevertheless I believe that at the end of the century the use of words will have altered so much that one will be able to speak of machines thinking without expecting to be contradicted."[10] Though this second prediction, about the change in our culture and the meaning of words, is less noted, it was prescient. It is through such changes in language—in how we speak about thinking, about machines wanting and liking things—that our culture and ethics evolve.

About fifteen years after "Computing Machinery and Intelligence" was published, Joseph Weizenbaum created the first program capable of carrying on such a text conversation. He named this program ELIZA, after the character in George Bernard Shaw's play *Pygmalion* who "learns to speak increasingly well."[11] Weizenbaum's research goal was to interact with computers using natural language; with this project he sought to show that a simple sentence-parsing program with some semantic heuristics could carry on a coherent conversation. ELIZA was able to find the topic of a sentence and had rules for forming a response, but had no contextual information about the world.

It was an approach quite different from what Turing envisioned. Turing's belief in the significance of carrying on a humanlike conversation was not as shallow an assumption as it seems now. He described a potentially winning machine as having processing power equivalent to the human brain (though he quite underestimated the human brain's complexity and power); it would initially be programmed to simulate an infant and would then be taught, much as a child is. Turing's views about the brain, learning, and children were remarkably naive. But the key point is that he believed that a machine that would pass his test would be one that was imbued with a mind analogous to that of humans, able to learn and to reason. Furthermore, though Turing remained adamant that we rely solely on external behavior in judging what is thinking, he outlined

[8] As a philosophical article, it is odd. It has pages of discussion about the nature of a digital computer but the central argument, that the Imitation Game is a satisfactory substitution for the question of whether machines can think, is rather glossed over.

[9] Specifically, that they would be able to "play the imitation game so well that an average interrogator will not have more than 70 per cent, chance of making the right identification after five minutes of questioning."

[10] Turing, "Computing Machinery and Intelligence," 442.

[11] Joseph Weizenbaum, "Contextual Understanding by Computers." *Communications of the ACM* 10, no. 8 (1967): 474.

the possibility of a state change, analogous to the critical mass of an atomic reaction that would mark a qualitative leap in mental ability and creativity.

ELIZA succeeded in sustaining conversation not through sophisticated technology but through, somewhat inadvertently, exploiting the way people make sense of each other. ELIZA was designed to respond based on scripts that would encode conversational rules for different roles. The first and by far most famous script Weizenbaum made for ELIZA was DOCTOR, modeled after a "Rogerian psychologist." His choice of this therapeutic framework was pragmatic: "the psychiatric interview is one of the few examples of categorized dyadic natural language communication in which one of the participating pair is free to assume the pose of knowing almost nothing of the real world."[12]

People were entranced with the computational "therapist." Even Weizenbaum's secretary, who knew the scope and point of the work, said upon trying it out that she wanted to chat with it further-in private.[13] Others took seriously the notion of the computational chat-bot as therapist, one that would be available to all, inexpensive and tireless.[14] At first Weizenbaum assumed this enthusiasm, which he judged to be misplaced, was due to the novelty of the interaction; future iterations should and would be designed to eliminate the "illusion of understanding."[15]

Weizenbaum's responses over the years show his growing alarm at this response. The quick willingness to accept a text-parsing program as an entity worthy of relating to, a repository for one's confidences, became to him an indicator of a deeply disturbing lack of concern about the humanity of the other—a lack of empathy and of even any interest in the mind and soul of the other. Weizenbaum had come to America fleeing Hitler's Europe and knew vividly and with horror the devastating effects of dehumanizing other people. He spent much of the rest of his career warning about the dangers computation posed to society.

Turing argued that we need to accept intelligent behavior (which he had redefined as the ability to convincingly imitate a human in a text conversation) as sufficient evidence of machine thinking. Fifteen years later, Weizenbaum's ELIZA, a clearly nonthinking, sentence-parsing chat-bot, posed a counterexample by demonstrating how easily the illusion of intelligence can be made. Dismayed by people's enthusiastic embrace of ELIZA's therapeutic potential (and computers in general), Weizenbaum came to believe that the willingness to accept machines in such roles was a significant threat to humane society. These positions, taken in the earliest years of AI research, delineate the big ethical questions surrounding artificial entities and provide the starting point for our analysis.

[12] Weizenbaum, "Contextual Understanding by Computers."

[13] Weizenbaum, "Contextual Understanding by Computers."

[14] Kenneth M. Colby, James B. Watt, and John P. Gilbert, "A Computer Method of Psychotherapy: Preliminary Communication," *Journal of Nervous and Mental Disease* 142, no. 2 (1966): 148–52.

[15] Joseph Weizenbaum, "ELIZA—a Computer Program for the Study of Natural Language Communication between Man and Machine," *Communications of the ACM* 9, no. 1 (1966): 43.

Where Are We Now?

Turing's prediction—that in limited conversations, machines would be indistinguishable from humans—was off by a few years. In 2000, there were no computers that were able to consistently pass as human after five minutes of text-based interaction. But a couple of decades later his prophesy has, effectively, come true.

In the narrow sense, computers have not "passed the Turing Test." There is an annual competition, the Loebner Prize, that takes Turing's Imitation Game suggestion literally, pitting a panel of judges against chat programs and hidden human typists. It has been widely criticized for encouraging programs that use tricks, such as simulated typing errors, to fool the judges, instead of advancing the goal of making more intelligent machines. Even so, while several have fooled judges during extended conversation, none has yet won the prize.

More significantly, we now interact with artificial entities in daily life, often without realizing they are not human. In 1950, when Turing proposed the Imitation Game, it was a stretch to think up a plausible scenario in which people would communicate via text with strangers of unknown and possibly fictitious identity. With the advent of the internet, this scenario has become commonplace.

In the mid-1990s, someone named Serdar Argic started inflaming the already heated Usenet arguments about the Armenian genocide by relentlessly posting hateful rants accusing the Armenians of massacring Turks. People wrote impassioned rebuttals to his screeds, thus making them even more disruptive by sidetracking any constructive discussion. Only after much anger and confusion did people realize that Argic was not a real person, but a program designed to intervene in any discussion that mentioned Armenia or Turkey, including Thanksgiving recipe posts. This was one of the first bots to deliberately fool people in a public setting.[16]

Chat-bots have since then become cleverer—and ubiquitous. They are tireless customer service agents, answering questions about ingredients, store hours, and mysterious error codes at any time of day or night. They are participants in online games, appearing as opponents, teammates, and incidental characters. They are the beautiful eager women in online dating sites who are always up for trying new things. Some are upfront about being software entities, but many attempt to pass as human.

An estimated 10–15 percent of users on the popular and influential social media site Twitter are bots. Some are useful: openly nonhuman programs that disseminate news, jokes, alerts, etc. But others masquerade as human users, seldom benevolently. They may be followers for hire, inflating their clients' apparent popularity. They may post vacation shots from sponsored villas, name-dropping restaurants, snacks, and songs, programmed to incessantly instigate flashes of envy and desire. Or they may be powerful purveyors of propaganda, chiming into political discussions, tirelessly hawking talking

[16] Judith Donath, *The Social Machine: Designs for Living Online* (Cambridge, MA: MIT Press, 2014).

points, slogans, and manufactured rumors. Bots thrive here in part because Twitter limits posts to 140 characters; non sequiturs, rather than back-and-forth discussions, characterize many interactions. Devising a program to mimic this style is much easier than creating one that must carry out an extended and coherent conversation.

Not all of today's artificial entities are online: we are increasingly surrounded by a growing population of social robots—autonomous, sentient-seeming objects. At home, we chat with friendly devices that fetch us the news, order us dinner, and ask politely about our day. We may have a robotic pet or coworker. There are robot receptionists who welcome guests in tech-forward hotels and robot orderlies who glide quietly into hospital rooms. Social robots are marketed as "friends" and "your next family member" who "can't wait to meet you."

No contemporary or readily foreseeable artificial entity is actually conscious or even primitively sentient, but our intuitive response to them is the opposite. They seem very much alert and aware. Our tendency to anthropomorphize contributes to this illusion. Yet when we see volition and intent in inanimate objects such as cars, trees, or dolls, we recognize that we ourselves are the source of its imagined vitality. With artificial entities, the object itself behaves in ways that strongly suggest a sentient experience lies within.

The ambiguity of their identity—machine or new form of thinking being—is no accident. Like the chat-bots that score highly in the Loebner Prize competition by making spelling mistakes, social robots are often made to mimic human habits such as pausing or looking away as if thinking; these easy-to-implement tricks provide a convincing illusion of sentience. Many are designed with simple, round childlike curves—features that elicit nurturance, indulgence, and trust[17], while also keeping our expectations of their abilities low. Their gendered voices and linguistic insinuation of self-conscious thought ("I'd like to help you") give the impression that one is speaking to an aware and sentient being.[18] As Turing predicted, our use of language has changed: we casually speak of these entities wanting, thinking, and liking.

Ethics of Our Relationship with the Seemingly Sentient

What are the ethical issues involved in our interaction with artificial entities? One set of issues concerns our responsibilities toward them—how we should treat them. The ethical framework I will use here is based on Peter Singer's utilitarian applied ethics;[19] his sentience-focused approach to assessing responsibilities toward nonhumans makes it

[17] Leslie Zebrowitz, *Reading Faces* (Boulder, CO: Westview Press, 1997).

[18] Friederike Eyssel, et al., "'If You Sound Like Me, You Must Be More Human': On the Interplay of Robot and User Features on Human-Robot Acceptance and Anthropomorphism." Paper presented at the 2012 7th ACM/IEEE International Conference on Human-Robot Interaction (HRI), 125–6.

[19] Singer, *Practical Ethics*.

especially relevant for thinking about artificial entities.[20] The key question here, however, is not how our treatment affects them, but what it does to us.

We noted earlier that our ethical responsibilities are to sentient beings: if something or someone has the capacity to feel, we need to take their preferences into consideration. To things that are not sentient—rocks, bacteria, dolls, robots—we have no direct moral obligation, that is, none that arises from their individual standing as a being with moral claims or rights. Since they do not experience anything, they cannot feel harmed by any action.

Though we do not have direct moral obligations to nonsentient entities that does not mean we have no obligations toward them. Nonconscious entities have what are called "indirect rights." These are rights that come from their relationship to a being that does have ethical standing; because harming the nonconscious entity would harm the being with ethical standing, it should therefore should be avoided. You adore your robot, and so I must treat it well because of your affection for it. It is wrong for me to harm something you value, not because of the intrinsic hurt to a thing (it has no feelings) but because you would be saddened by its loss.

Laws reflect a society's ethics, but they change slowly and are often more an indicator of the morals of its past. Indirect rights have been the primary source of protection that animals have had under American law: I cannot kick your dog, not because it would hurt your dog but because you would be upset (and it is your property). Indirect rights are often weak. In the moral calculus required to balance numerous competing preferences and rights, they can be readily eclipsed. Protection based on human preference disappears in the face of competing human interests—thus we have factory farms, sport hunting, etc.

Society changes. Laws protecting animals based on ethical reasoning that takes their experience into account—that recognizes their sentience—are becoming more common. The change is due both to (a) seeing sentience as the quality that defines whether one has direct moral claims and (b) recognizing that some animals are sentient. It is also part of a broader Western cultural shift to an increasingly inclusive view of who is a being with moral standing: it is not that long ago in the United States that women and slaves had mainly indirect rights. Advocates for animal rights posit that what they call "speciesism"—the belief that members of one species have superior moral standing on the basis of that membership—as the logical and moral equivalent of racism.

Some legal scholars have argued that such legal protection should extend to social robots:[21] "We may not want to be the kind of society that tolerates cruelty to an entity we

[20] The focus of this chapter is on Western society. See Frédéric Kaplan, "Who Is Afraid of the Humanoid? Investigating Cultural Differences in the Acceptance of Robots," *International journal of Humanoid Robotics* 1, no. 03 (2004): 465–80; and Jennifer Robertson, *Robo Sapiens Japanicus: Robots, Gender, Family, and the Japanese Nation* (Berkeley: University of California Press, 2017), for reactions to artificial entities in Japan.

[21] Kate Darling, "Extending Legal Protection to Social Robots: The Effects of Anthropomorphism, Empathy, and Violent Behavior towards Robotic Objects," in *Robot Law*, ed. A. M. Froomkin R. Calo, and I. Kerr. (Cheltenhem, UK: Edward Elgar, 2016), 213–34.

think of as quasi-human."[22] I argue that this movement toward more inclusive rights does not, and should not, apply to nonsentient artificial beings. The fundamental reason for extending moral rights to animals is recognition of their sentience—that they can experience suffering. It is a right inherent to them, regardless of whether a human observer, owner, or other interested party is aware of their pain.[23] The premise that sentience is the foundation of moral rights is important—extending these rights to nonsentient entities dilutes its meaning and significance.

That said, the compelling simulation of sentience exhibited by artificial entities can provide them with additional indirect moral claims, again stemming from considerations about a person's experience, not the entity's. Here the concern is that treating another cruelly brutalizes oneself. This principle is reflected in Jewish custom, which forbids sport hunting because it encourages cruelty, even if the animal is killed painlessly.[24] And Immanuel Kant, though he argued that animals have no "will" and thus no inherent rights, also wrote, "If he is not to stifle his own feelings, he must practice kindness towards animals, for he who is cruel to animals becomes hard also in his dealings with men."[25]

Behaving ethically often involves trade-offs between competing rights and principles, and even a seemingly simple injunction such as "do not treat sentient-seeming entities cruelly" can create dilemmas. The popular keychain pet toy, the Tamagotchi, provides a useful scenario. These are very simple artificial entities that nonetheless exert a powerful emotional pull.[26] The owner of a Tamagotchi must work at keeping it "alive," a task that entails pushing buttons on it at frequent but arbitrary times. Ignore it and it will cease to thrive and will eventually "die"; as with real pets, cruelty toward the Tamagotchi can take the form of neglect. Imagine now a family dinner. The grandmother is visiting, but a grandchild is continuously distracted, checking a Tamagotchi's status. Should the parents demand the child put the toy away and pay full attention to the (living, conscious, and closely related) grandparent present in the room, who would like their attention, but at the cost of allowing the Tamagotchi to possibly die? Or is nurturing the keychain pet useful training in responsible caring, so grandmother and virtual pet will need to share the child's divided attention?

The appeal of the simple Tamagotchi vividly demonstrates just how compelling and potentially manipulative an artificial entity can be. This raises concerns about prohibitions against mistreating them—and especially about encasing such prohibitions in law. The makers of an artificial entity can design it so that arbitrary events and conditions

[22] Ryan Calo, "Robotics and the Lessons of Cyberlaw," *California Law Review* (2015): 513–63.

[23] Darling points out that animal protection law seems to reflect the popular sentimental standing of particular animals, rather than the philosophically or biologically based concern with their sentience. In this chapter, our focus is on fundamental ethics—on getting the theory right in order to guide the practice.

[24] Rabbi Dr. Asher Meir, "Judaism and Hunting," *Jewish Ethicist*, https://www.ou.org/torah/machshava/jewish-ethicist/judaism_and_hunting/.

[25] Immanuel Kant, *Lectures on Ethics*, trans. Louis Infield. (New York: Harper & Row, 1963), 240.

[26] Frédéric Kaplan, "Free Creatures: The Role of Uselessness in the Design of Artificial Pets" (paper presented at the 1st Edutainment Robotics Workshop, Sankt Augustin, Germany, 2000), 45–7.

burdens, to test medicines, and to entertain us—and the relief from responsibility that comes with insisting, even in the face of vivid contrary evidence, that they are incapable of suffering.

Our dystopian predictions of what a powerful and conscious machine would do are not based on projection from the technology or even from biology. They seem, instead, like the nightmares of a guilty conscience. The ethical challenge is to use this existential guilt to change. Can we treat the other beings we live with on Earth as we would want conscious, super-powerful artificial entities to treat us?

Bibliography

Broom, Donald M. *Sentience and Animal Welfare*. Boston: CABI, 2014.

Calo, Ryan. "Robotics and the Lessons of Cyberlaw." *California Law Review* (2015): 513–63.

DePaulo, Bella M., Deborah A. Kashy, Susan E. Kirkendol, Melissa M. Wyer, and Jennifer A. Epstein. "Lying in Everyday Life." *Journal of Personality and Social Psychology* 70, no. 5 (1996): 979.

Donath, Judith. "The Robot Dog Fetches for Whom?" In *A Networked Self and Human Augmentics, Artificial Intelligence, Sentience*, edited by Zizi Papacharissi, 26–40. New York: Routledge, 2018.

Godfrey-Smith, Peter. *Other Minds: The Octopus and the Evolution of Intelligent Life*. London: William Collins, 2016.

Kaplan, Frédéric. "Who Is Afraid of the Humanoid? Investigating Cultural Differences in the Acceptance of Robots." *International Journal of Humanoid Robotics* 1, no. 3 (2004): 465–80.

Singer, Peter. *Practical Ethics*. Cambridge, UK: Cambridge University Press, 2011.

Turing, Alan. "Computing Machinery and Intelligence." *Mind* 49 (1950): 433–60.

Turkle, Sherry. "Authenticity in the Age of Digital Companions." *Interaction Studies* 8, no. 3 (2007): 501–17.

Weizenbaum, Joseph. *Computer Power and Human Reason*. San Francisco: W. H. Freeman, 1976.

Weizenbaum, Joseph. "Contextual Understanding by Computers." *Communications of the ACM* 10, no. 8 (1967): 474–80.

PART II

FRAMEWORKS AND MODES

socio-technical systems that utilize data-driven algorithms to classify, to make decisions, and to control complex systems, including the use of machine learning and large datasets to generate predictions about future behavior (hereafter "AI" systems")[2], may interfere with human rights. The recent Cambridge Analytica scandal revealed how unlawfully harvested Facebook data from millions of voters in the United Kingdom, the United States, and elsewhere enabled malign actors to engage in political micro-targeting through the use of AI-driven social media content distribution systems, thereby interfering with their right to free and fair elections and thus threatening the integrity of democratic processes. The increasing use of algorithmic decision-making (ADM) systems to inform custodial and other decisions within the criminal justice process may threaten several human rights, including the right to a fair trial, the presumption of innocence, and the right to liberty and security. Systems of this kind are now used to inform, and often to automate, decisions about an individual's eligibility and entitlement to various benefits and opportunities, including housing, social security, finance, employment and other life-affecting opportunities, potentially interfering with rights of due process and rights to freedom from unfair or unlawful discrimination.[3] Because these systems have the capacity to operate both automatically and at scale, their capacity to affect thousands if not millions of people at a stroke can now occur at orders of magnitude and speeds not previously possible.[4]

This chapter has two overarching aims. Firstly, we argue that the international human rights framework provides the most promising set of standards for ensuring that AI systems are ethical in their design, development, and deployment. Secondly, we sketch the basic contours of a comprehensive governance framework, which we refer to as a human rights–centered design, deliberation, and oversight approach for ensuring that AI can be relied upon to operate in ways that will not violate human rights.

Four features of ongoing discussions provide important contexts for our argument. First, the rubric of "AI ethics" is now used to encapsulate a multiplicity of value-based, societal concerns associated with the use of AI applications across an increasingly extensive and diverse range of social and economic activities. Second, there is a notable

Profiled: Cogitas Ergo Sum, ed. Bayamlioglu, Irina Baraliuc, Liisa Janssens, and Mireille Hildebrandt (Amsterdam: Amsterdam University Press, 2018).

[2] For more detail, see European Commission, "Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions on Artificial Intelligence for Europe," 2018, https://ec.europa.eu/digital-single-market/en/news/communication-artificial-intelligence-europe (accessed March 17, 2020).

[3] B. Wagner, *Study on the Human Rights Dimensions of Automated Data Processing Techniques (in Particular Algorithms) and Possible Regulatory Implications*, Council of Europe, Committee of experts on internet intermediaries (MSI-NET), 2017, https://rm.coe.int/study-hr-dimension-of-automated-data-processing-incl-algorithms/168075b94a (accessed June 11, 2018).

[4] Karen Yeung, *A Study of the Implications of Advanced Digital Technologies (Including AI Systems) for the Concept of Responsibility within a Human Rights Framework*, Council of Europe MSI-AUT committee study, 2019, DGI(2019)05, https://rm.coe.int/a-study-of-the-implications-of-advanced-digital-technologies-including/168096bdab (accessed December 9, 2019).

CHAPTER 24

A HUMAN-CENTERED APPROACH TO AI ETHICS

A Perspective from Cognitive Science

RON CHRISLEY

The increasing role of artificial intelligence (AI) and machine learning technology in our lives has raised an enormous number and variety of ethical challenges, as can be seen in the diverse topics covered in this volume. In addition, there are the ethical challenges yet to come, ones that we cannot currently anticipate. We can try to respond to this vast array of challenges individually, in an ad hoc manner, but in the long run, a more principled, structured response is likely to be of more guidance. In this chapter I propose responses to some particular questions concerning the ethics of AI, responses that share a unifying perspective: a human-centered approach. The hope is that, beyond offering solutions to the particular problems considered here, these responses can be of more general interest by illuminating enough of their shared, human-centered perspective to facilitate like-minded responses to any number of current and future ethical challenges involving AI.

More will be said about what the human-centered approach to AI/robot ethics amounts to, but an important consequence of it, and the central claim of this chapter, is this: when making ethical judgments in this area, we should resist the temptation to see robots as ethical agents or patients. For the foreseeable future, more ethical hazard follows from seeing humans and robots as ethically analogous than follows from seeing them as ethically distinct kinds. Much of what I say in what follows is meant to support this claim, to identify some instances of current practice that fail to heed the warnings of the claim and to suggest ways of avoiding the anthropomorphic error the claim identifies, while still minimizing the likelihood of certain ethically adverse outcomes involving robots and AI in general.

This central claim can seem at odds with an otherwise attractive naturalism about ethics, mind, and what it is to be human. My adoption of a human-centered approach to the ethics of AI arises out of my lifelong interest in cognitive science. Cognitive science is the interdisciplinary search for an understanding of how mentality in general (not just cognition) can be part of the natural world, and the use of that understanding to provide explanations of mental phenomena and the behavior of systems with minds. One might think that this naturalism (particularly in the mechanistic, functionalist, physicalist form that many traditional cognitive scientists embrace, even if only implicitly) encourages us to see ourselves as glorified robots, a rough equation that would either support the extension of the concepts of ethical agent and patient to suitably programmed robots and AI systems, or encourage ethical nihilism for both humans and robots. Contrary to this, I believe that seeing humans as part of the natural world does not undermine our understanding of what makes humans ethically different from robots (or nonhuman animals); rather, it gives that understanding scientific plausibility and conceptual clarity. It is only by properly considering our place in the natural world that we can see the true, nondualist, reasons why it is correct to see us, but not robots (at least for the forseeable future), as ethical beings. Nevertheless, the theories and methods of cognitive science will largely remain in the background of this chapter, with the focus instead being on the human-centered approach they support.

Putting Robots in Their Place

Just what do I mean by a human-centered approach? We'll be better equipped to answer that question in full after we have a few instances of it from which to generalize, but a few things can be said at the outset to give an initial idea of what the approach is—and what it is not.

The human-centered approach to AI ethics I am advocating here has two key aspects:

1. An emphasis on human welfare.
2. An emphasis on human responsibility.

The first aspect is in contrast with approaches to AI ethics that take seriously ethical obligations concerning the purported welfare of artificial agents. Such approaches focus on questions such as:

- Can robots feel pain?
- Can they suffer?
- If so, what are our obligations, if any, for reducing robot pain and suffering at the expense of increasing human pain and suffering? At the expense of increasing animal pain and suffering?

Similarly, the second aspect of the human-centered approach is in contrast with approaches that focus on questions such as:

Index

Note: Figures are indicated by an *f* following the page number.