Dædalus

Journal of the American Academy of Arts & Sciences

Spring 2022

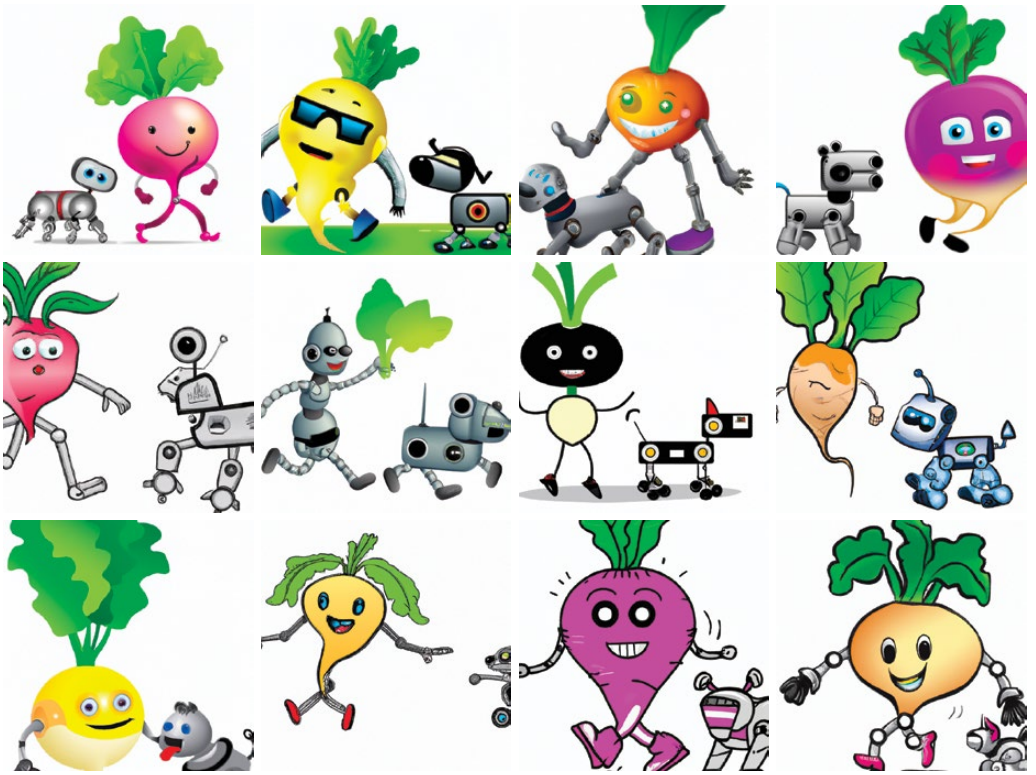
AI & Society



James Manyika, guest editor

with Nigel Shadbolt · Stuart Russell
Jeffrey Dean · Kevin Scott · Li Fei-Fei
Ranjay Krishna · Daniela Rus · Kobi Gal
Barbara J. Grosz · Christopher D. Manning
Yejin Choi · Emira Murati · Tobias Rees
Blaise Agüera y Arcas · Michele Elam
Iason Gabriel · John Tasioulas
Michael Spence · Laura D. Tyson
John Zysman · Erik Brynjolfsson
Eric Schmidt · Ash Carter · Cynthia Dwork
Martha Minow · Sonia K. Katyal
Mariano-Florentino Cuéllar · Aziz Z. Huq
Diane Coyle · Helen Margetts

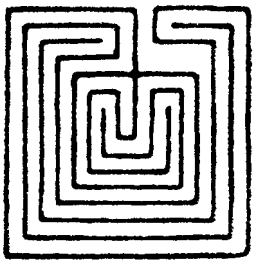
Prompt: *An illustration of a happy turnip walking a dog-robot.*
Generations:



Prompt: *A robot testing a human if the human is a real human.*
Generations:



Dædalus



Journal of the American Academy of Arts & Sciences

"AI & Society"

Volume 151, Number 2; Spring 2022

James Manyika, Guest Editor

Phyllis S. Bendell, Managing Editor and Director of Publications

Peter Walton, Associate Editor

Heather M. Struntz, Assistant Editor

© 2022 by the American Academy of Arts & Sciences.

Dædalus is published under a Creative Commons Attribution-NonCommercial 4.0 International license (CC BY-NC 4.0). For allowed uses under this license, please visit creativecommons.org/licenses/by-nc/4.0/.

Dædalus publishes by invitation only and assumes no responsibility for unsolicited manuscripts. The views expressed are those of the author(s) of each essay, and not necessarily of the American Academy of Arts & Sciences.

Library of Congress Catalog No. 12-30299. ISSN 0011-5266; E-ISSN 1548-6192.
An electronic full-text version of *Dædalus* is available from amacad.org/daedalus.
Please direct questions or comments to daedalus@amacad.org.

The images appearing on the inside covers of this issue of *Dædalus* were generated by a state-of-the-art version of GPT-3 (Generative Pre-trained Transformer 3) that builds on the approaches used in DALL·E and GLIDE. Given a text prompt that it has not been trained for or previously exposed to, the model generates original images based on its understanding of the different elements in the prompt and how they relate to each other. Given the same prompt repeatedly, the model produces novel responses. The groups of images here consist of unique outputs to the same prompts, which appear above each set. For more on GPT-3, see the afterword to this volume.

Contents

5 **Getting AI Right: Introductory Notes on AI & Society**
James Manyika

On beginnings & progress

28 **"From So Simple a Beginning": Species of Artificial Intelligence**
Nigel Shadbolt

43 **If We Succeed**
Stuart Russell

On building blocks, systems & applications

58 **A Golden Decade of Deep Learning: Computing Systems & Applications**
Jeffrey Dean

75 **I Do Not Think It Means What You Think It Means: Artificial Intelligence, Cognitive Work & Scale**
Kevin Scott

On machine vision, robots & agents

85 **Searching for Computer Vision North Stars**
Li Fei-Fei & Ranjay Krishna

100 **The Machines from Our Future**
Daniela Rus

114 **Multi-Agent Systems: Technical & Ethical Challenges of Functioning in a Mixed Group**
Kobi Gal & Barbara J. Grosz

On language & reasoning

127 **Human Language Understanding & Reasoning**
Christopher D. Manning

139 The Curious Case of Commonsense Intelligence
Yejin Choi

156 Language & Coding Creativity
Ermira Murati

On philosophical laboratories & mirrors

168 Non-Human Words: On GPT-3 as a Philosophical Laboratory
Tobias Rees

183 Do Large Language Models Understand Us?
Blaise Agüera y Arcas

198 Signs Taken for Wonders: AI, Art & the Matter of Race
Michele Elam

On inequality, justice & ethics

218 Toward a Theory of Justice for Artificial Intelligence
Iason Gabriel

232 Artificial Intelligence, Humanistic Ethics
John Tasioulas

On the economy & future of work

244 Automation, Augmentation, Value Creation & the Distribution of Income & Wealth
Michael Spence

256 Automation, AI & Work
Laura D. Tyson & John Zysman

272 The Turing Trap: The Promise & Peril of Human-Like Artificial Intelligence
Erik Brynjolfsson

On great power competition & national security

288 **AI, Great Power Competition & National Security**
Eric Schmidt

299 **The Moral Dimension of AI-Assisted Decision-Making: Some Practical Perspectives from the Front Lines**
Ash Carter

On the law & public trust

309 **Distrust of Artificial Intelligence: Sources & Responses from Computer Science & Law**
Cynthia Dwork & Martha Minow

322 **Democracy & Distrust in an Era of Artificial Intelligence**
Sonia K. Katyal

335 **Artificially Intelligent Regulation**
Mariano-Florentino Cuéllar & Aziz Z. Huq

On seeing rooms & governance

348 **Socializing Data**
Diane Coyle

360 **Rethinking AI for Good Governance**
Helen Margetts

372 **Afterword: Some Illustrations**
James Manyika

Getting AI Right: Introductory Notes on AI & Society

James Manyika

NATHAN: Do you know what the Turing Test is?

CALEB: . . . Yeah. I know what the Turing Test is. It's when a human interacts with a computer. And if the human doesn't know they're interacting with a computer, the test is passed.

NATHAN: And what does a pass tell us?

CALEB: That the computer has artificial intelligence. . . .

NATHAN: You got it. Because if that test is passed, you are dead center of the single greatest scientific event in the history of man.

CALEB: If you've created a conscious machine, it's not the history of man. It's the history of gods.

This dialogue is from an early scene in the 2014 film *Ex Machina*, in which Nathan has invited Caleb to determine whether Nathan has succeeded in creating artificial intelligence.[1] The achievement of powerful artificial general intelligence has long held a grip on our imagination not only for its exciting as well as worrisome possibilities, but also for its suggestion of a new, uncharted era for humanity. In opening his 2021 BBC Reith Lectures, titled "Living with Artificial Intelligence," Stuart Russell states that "the eventual emergence of general-purpose artificial intelligence [will be] the biggest event in human history."[2]

Over the last decade, a rapid succession of impressive results has brought wider public attention to the possibilities of powerful artificial intelligence. In machine vision, researchers demonstrated systems that could recognize objects as well as, if not better than, humans in some situations. Then came the games. Complex games of strategy have long been associated with superior intelligence, and so when AI systems beat the best human players at chess, Atari games, Go, shogi, StarCraft, and Dota, the world took notice. It was not just that AIs beat humans (although that was astounding when it first happened), but the escalating progression of how they did it: initially by learning from expert human play, then from self-play, then by teaching themselves the principles of the games from the ground up, eventually yielding single systems that could learn, play, and win at

© 2022 by James Manyika
Published under a Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) license
https://doi.org/10.1162/DAED_e_01897

several structurally different games, hinting at the possibility of generally intelligent systems.[3]

Speech recognition and natural language processing have also seen rapid and headline-grabbing advances. Most impressive has been the emergence recently of large language models capable of generating human-like outputs. Progress in language is of particular significance given the role language has always played in human notions of intelligence, reasoning, and understanding. While the advances mentioned thus far may seem abstract, those in driverless cars and robots have been more tangible given their embodied and often biomorphic forms. Demonstrations of such embodied systems exhibiting increasingly complex and autonomous behaviors in our physical world have captured public attention.

Also in the headlines have been results in various branches of science in which AI and its related techniques have been used as tools to advance research from materials and environmental sciences to high energy physics and astronomy.[4] A few highlights, such as the spectacular results on the fifty-year-old protein-folding problem by AlphaFold, suggest the possibility that AI could soon help tackle science's hardest problems, such as in health and the life sciences.[5]

While the headlines tend to feature results and demonstrations of a future to come, AI and its associated technologies are already here and pervade our daily lives more than many realize. Examples include recommendation systems, search, language translators – now covering more than one hundred languages – facial recognition, speech to text (and back), digital assistants, chatbots for customer service, fraud detection, decision support systems, energy management systems, and tools for scientific research, to name a few. In all these examples and others, AI-related techniques have become components of other software and hardware systems as methods for learning from and incorporating messy real-world inputs into inferences, predictions, and, in some cases, actions. As director of the Future of Humanity Institute at the University of Oxford, Nick Bostrom noted back in 2006, "A lot of cutting-edge AI has filtered into general applications, often without being called AI because once something becomes useful enough and common enough it's not labeled AI anymore."[6]

As the scope, use, and usefulness of these systems have grown for individual users, researchers in various fields, companies and other types of organizations, and governments, so too have concerns when the systems have not worked well (such as bias in facial recognition systems), or have been misused (as in deepfakes), or have resulted in harms to some (in predicting crime, for example), or have been associated with accidents (such as fatalities from self-driving cars).[7]

Dædalus last devoted a volume to the topic of artificial intelligence in 1988, with contributions from several of the founders of the field, among others. Much of that issue was concerned with questions of whether research in AI was making progress, of whether AI was at a turning point, and of its foundations, mathemati-

cal, technical, and philosophical – with much disagreement. However, in that volume there was also a recognition, or perhaps a rediscovery, of an alternative path toward AI – the connectionist learning approach and the notion of neural nets – and a burgeoning optimism for this approach's potential. Since the 1960s, the learning approach had been relegated to the fringes in favor of the symbolic formalism for representing the world, our knowledge of it, and how machines can reason about it. Yet no essay captured some of the mood at the time better than Hilary Putnam's "Much Ado About Not Very Much." Putnam questioned the *Dædalus* issue itself: "Why a whole issue of *Dædalus*? Why don't we wait until AI achieves something and then have an issue?" He concluded:

> Perhaps the optimistic view is right, but I do not see anyone on the scene, in either artificial intelligence or inductive logic, who has any interesting ideas about how the topic-neutral [general] learning strategy works. When someone does appear with such an idea, that will be time for *Dædalus* to publish an issue on AI.[8]

This volume of *Dædalus* is indeed the first since 1988 to be devoted to artificial intelligence. This volume does not rehash the same debates; much else has happened since, mostly as a result of the success of the machine learning approach that was being rediscovered and reimagined, as discussed in the 1988 volume. This issue aims to capture where we are in AI's development and how its growing uses impact society. The themes and concerns herein are colored by my own involvement with AI. Besides the television, films, and books that I grew up with, my interest in AI began in earnest in 1989 when, as an undergraduate at the University of Zimbabwe, I undertook a research project to model and train a neural network.[9] I went on to do research on AI and robotics at Oxford. Over the years, I have been involved with researchers in academia and labs developing AI systems, studying AI's impact on the economy, tracking AI's progress, and working with others in business, policy, and labor grappling with its opportunities and challenges for society.[10]

The authors of the twenty-five essays in this volume range from AI scientists and technologists at the frontier of many of AI's developments to social scientists at the forefront of analyzing AI's impacts on society. The volume is organized into ten sections. Half of the sections are focused on AI's development, the other half on its intersections with various aspects of society. In addition to the diversity in their topics, expertise, and vantage points, the authors bring a range of views on the possibilities, benefits, and concerns for society. I am grateful to the authors for accepting my invitation to write these essays.

Before proceeding further, it may be useful to say what we mean by artificial intelligence. The headlines and increasing pervasiveness of AI and its associated technologies have led to some conflation and confusion about

what exactly counts as AI. This has not been helped by the current trend – among researchers in science and the humanities, startups, established companies, and even governments – to associate anything involving not only machine learning, but data science, algorithms, robots, and automation of all sorts with AI. This could simply reflect the hype now associated with AI, but it could also be an acknowledgment of the success of the current wave of AI and its related techniques and their wide-ranging use and usefulness. I think both are true ; but it has not always been like this. In the period now referred to as the AI winter, during which progress in AI did not live up to expectations, there was a reticence to associate most of what we now call AI with AI.

Two types of definitions are typically given for AI. The first are those that suggest that it is the ability to artificially do what intelligent beings, usually human, can do. For example, artificial intelligence is :

> the ability of a digital computer or computer-controlled robot to perform tasks commonly associated with intelligent beings.[11]

The human abilities invoked in such definitions include visual perception, speech recognition, the capacity to reason, solve problems, discover meaning, generalize, and learn from experience. Definitions of this type are considered by some to be limiting in their human-centricity as to what counts as intelligence and in the benchmarks for success they set for the development of AI (more on this later). The second type of definitions try to be free of human-centricity and define an intelligent agent or system, whatever its origin, makeup, or method, as :

> Any system that perceives its environment and takes actions that maximize its chance of achieving its goals.[12]

This type of definition also suggests the pursuit of goals, which could be given to the system, self-generated, or learned.[13] That both types of definitions are employed throughout this volume yields insights of its own.

These definitional distinctions notwithstanding, the term AI, much to the chagrin of some in the field, has come to be what cognitive and computer scientist Marvin Minsky called a "suitcase word."[14] It is packed variously, depending on who you ask, with approaches for achieving intelligence, including those based on logic, probability, information and control theory, neural networks, and various other learning, inference, and planning methods, as well as their instantiations in software, hardware, and, in the case of embodied intelligence, systems that can perceive, move, and manipulate objects.

Three questions cut through the discussions in this volume : 1) Where are we in AI's development ? 2) What opportunities and challenges does AI pose for society ? 3) How much about AI is really about us ?

Where are we in AI's development?

Notions of intelligent machines date all the way back to antiquity.[15] Philosophers, too, among them Hobbes, Leibnitz, and Descartes, have been dreaming about AI for a long time; Daniel Dennett suggests that Descartes may have even anticipated the Turing Test.[16] The idea of computation-based machine intelligence traces to Alan Turing's invention of the universal Turing machine in the 1930s, and to the ideas of several of his contemporaries in the mid-twentieth century. But the birth of artificial intelligence as we know it and the use of the term is generally attributed to the now famed Dartmouth summer workshop of 1956. The workshop was the result of a proposal for a two-month summer project by John McCarthy, Marvin Minsky, Nathaniel Rochester, and Claude Shannon whereby "An attempt will be made to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves."[17]

In their respective contributions to this volume, "From So Simple a Beginning: Species of Artificial Intelligence" and "If We Succeed," and in different but complementary ways, Nigel Shadbolt and Stuart Russell chart the key ideas and developments in AI, its periods of excitement as well as the aforementioned AI winters. The current AI spring has been underway since the 1990s, with headline-grabbing breakthroughs appearing in rapid succession over the last ten years or so: a period that Jeffrey Dean describes in the title of his essay as a "golden decade," not only for the pace of AI development but also its use in a wide range of sectors of society, as well as areas of scientific research.[18] This period is best characterized by the approach to achieve artificial intelligence through learning from experience, and by the success of neural networks, deep learning, and reinforcement learning, together with methods from probability theory, as ways for machines to learn.[19]

A brief history may be useful here: In the 1950s, there were two dominant visions of how to achieve machine intelligence. One vision was to use computers to create a logic and symbolic representation of the world and our knowledge of it and, from there, create systems that could reason about the world, thus exhibiting intelligence akin to the mind. This vision was most espoused by Allen Newell and Hebert Simon, along with Marvin Minsky and others. Closely associated with it was the "heuristic search" approach that supposed intelligence was essentially a problem of exploring a space of possibilities for answers. The second vision was inspired by the brain, rather than the mind, and sought to achieve intelligence by learning. In what became known as the connectionist approach, units called perceptrons were connected in ways inspired by the connection of neurons in the brain. At the time, this approach was most associated with Frank Rosenblatt. While there was initial excitement about both visions, the first came to dominate, and did so for decades, with some successes, including so-called expert systems.

Not only did this approach benefit from championing by its advocates and plentiful funding, it came with the suggested weight of a long intellectual tradition – exemplified by Descartes, Boole, Frege, Russell, and Church, among others – that sought to manipulate symbols and to formalize and axiomatize knowledge and reasoning. It was only in the late 1980s that interest began to grow again in the second vision, largely through the work of David Rumelhart, Geoffrey Hinton, James McClelland, and others. The history of these two visions and the associated philosophical ideas are discussed in Hubert Dreyfus and Stuart Dreyfus's 1988 *Dædalus* essay "Making a Mind Versus Modeling the Brain : Artificial Intelligence Back at a Branchpoint."[20] Since then, the approach to intelligence based on learning, the use of statistical methods, back-propagation, and training (supervised and unsupervised) has come to characterize the current dominant approach.

Kevin Scott, in his essay "I Do Not Think It Means What You Think It Means : Artificial Intelligence, Cognitive Work & Scale," reminds us of the work of Ray Solomonoff and others linking information and probability theory with the idea of machines that can not only learn, but compress and potentially generalize what they learn, and the emerging realization of this in the systems now being built and those to come. The success of the machine learning approach has benefited from the boon in the availability of data to train the algorithms thanks to the growth in the use of the Internet and other applications and services. In research, the data explosion has been the result of new scientific instruments and observation platforms and data-generating breakthroughs, for example, in astronomy and in genomics. Equally important has been the co-evolution of the software and hardware used, especially chip architectures better suited to the parallel computations involved in data- and compute-intensive neural networks and other machine learning approaches, as Dean discusses.

Several authors delve into progress in key subfields of AI.[21] In their essay, "Searching for Computer Vision North Stars," Fei-Fei Li and Ranjay Krishna chart developments in machine vision and the creation of standard data sets such as ImageNet that could be used for benchmarking performance. In their respective essays "Human Language Understanding & Reasoning" and "The Curious Case of Commonsense Intelligence," Chris Manning and Yejin Choi discuss different eras and ideas in natural language processing, including the recent emergence of large language models comprising hundreds of billions of parameters and that use transformer architectures and self-supervised learning on vast amounts of data.[22] The resulting pretrained models are impressive in their capacity to take natural language prompts for which they have not been trained specifically and generate human-like outputs, not only in natural language, but also images, software code, and more, as Mira Murati discusses and illustrates in "Language & Coding Creativity." Some have started to refer to these large language models as foundational models in that once they are trained, they are adaptable to a wide range of tasks and outputs.[23] But

despite their unexpected performance, these large language models are still early in their development and have many shortcomings and limitations that are highlighted in this volume and elsewhere, including by some of their developers.[24]

In "The Machines from Our Future," Daniela Rus discusses the progress in robotic systems, including advances in the underlying technologies, as well as in their integrated design that enables them to operate in the physical world. She highlights the limitations in the "industrial" approaches used thus far and suggests new ways of conceptualizing robots that draw on insights from biological systems. In robotics, as in AI more generally, there has always been a tension as to whether to copy or simply draw inspiration from how humans and other biological organisms achieve intelligent behavior. Elsewhere, AI researcher Demis Hassabis and colleagues have explored how neuroscience and AI learn from and inspire each other, although so far more in one direction than the other, as Alexis Baria and Keith Cross have suggested.[25]

Despite the success of the current approaches to AI, there are still many shortcomings and limitations, as well as conceptually hard problems in AI.[26] It is useful to distinguish on one hand problematic shortcomings, such as when AI does not perform as intended or safely, or produces biased or toxic outputs that can lead to harm, or when it impinges on privacy, or generates false information about the world, or when it has characteristics such as lack of explainability, all of which can lead to a loss of public trust. These shortcomings have rightly captured the attention of the wider public and regulatory bodies, as well as researchers, among whom there is an increased focus on technical AI and ethics issues.[27] In recent years, there has been a flurry of efforts to develop principles and approaches to responsible AI, as well as bodies involving industry and academia, such as the Partnership on AI, that aim to share best practices.[28] Another important shortcoming has been the significant lack of diversity–especially with respect to gender and race–in the people researching and developing AI in both industry and academia, as has been well documented in recent years.[29] This is an important gap in its own right, but also with respect to the characteristics of the resulting AI and, consequently, in its intersections with society more broadly.

On the other hand, there are limitations and hard problems associated with the things that AI is not yet capable of that, if solved, could lead to more powerful, more capable, or more general AI. In their Turing Lecture, deep learning pioneers Yoshua Bengio, Yann LeCun, and Geoffrey Hinton took stock of where deep learning stands and highlighted its current limitations, such as the difficulties with out-of-distribution generalization.[30] In the case of natural language processing, Manning and Choi highlight the hard challenges in reasoning and commonsense understanding, despite the surprising performance of large language models. Elsewhere, computational linguists Emily Bender and Alexander Koller have challenged the notion that large language models do anything resembling under-

standing, learning, or meaning.[31] In "Multi-Agent Systems: Technical & Ethical Challenges of Functioning in a Mixed Group," Kobi Gal and Barbara Grosz discuss the hard problems in multi-agent systems, highlighting the conceptual difficulties – such as how to reason about other agents, their belief systems, and intentionality – as well as ethical challenges in both cooperative and competitive settings, especially when the agents include both humans and machines. Elsewhere, Allan Dafoe and others provide a useful overview of the open problems in cooperative AI.[32] Indeed, there is a growing sense among many that we do not have adequate theories for the sociotechnical embedding of AI systems, especially as they become more capable and the scope of societal use expands.

And although AI and its related techniques are proving to be powerful tools for research in science, as examples in this volume and elsewhere illustrate – including recent examples in which embedded AI capabilities not only help evaluate results but also steer experiments by going beyond heuristics-based experimental design and become what some have termed "self-driving laboratories"[33] – getting AI to understand science and mathematics and to theorize and develop novel concepts remain grand challenges for AI.[34] Indeed the possibility that more powerful AI could lead to new discoveries in science, as well as enable game-changing progress in some of humanity's greatest challenges and opportunities, has long been a key motivation for many at the frontier of AI research to build more capable systems.

Beyond the particulars of each subfield of AI, the list of more general hard problems that continue to limit the possibility of more capable AI includes one-shot learning, cross-domain generalizations, causal reasoning, grounding, complexities of timescales and memory, and meta-cognition.[35] Consideration of these and other hard problems that could lead to more capable systems raises the question of whether current approaches – mostly characterized by deep learning, the building of larger and larger and more foundational and multimodal models, and reinforcement learning – are sufficient, or whether entirely different conceptual approaches are needed in addition, such as neuroscience-inspired cognitive agent approaches or semantic representations or reasoning based on logic and probability theory, to name a few. On whether and what kind of additional approaches might be needed, the AI community is divided, but many believe the current approaches[36] along with further evolution of compute and learning architectures have yet to reach their limits.[37]

The debate about the sufficiency of the current approaches is closely associated with the question of whether artificial general intelligence can be achieved, and if so, how and when. *Artificial general intelligence* (AGI) is defined in distinction to what is sometimes called *narrow AI*: that is, AI developed and fine-tuned for specific tasks and goals, such as playing chess. The development of AGI, on the other hand, aims for more powerful AI – at least as powerful as humans – that is generally applicable to any problem or situation and, in some conceptions, includes the capacity to evolve and improve itself, as well as set and evolve its own goals and

preferences. Though the question of whether, how, and when AGI will be achieved is a matter for debate, most agree that its achievement would have profound implications – beneficial and worrisome – for humanity, as is often depicted in popular books[38] and films such as *2001: A Space Odyssey* through *Terminator* and *The Matrix* to *Ex Machina* and *Her*. Whether it is imminent or not, there is growing agreement among many at the frontier of AI research that we should prepare for the possibility of powerful AGI with respect to safety and control, alignment and compatibility with humans, its governance and use, and the possibility that multiple varieties of AGI could emerge, and that we should factor these considerations into how we approach the development of AGI.

Most of the investment, research and development, and commercial activity in AI today is of the narrow AI variety and in its numerous forms: what Nigel Shadbolt terms the *speciation* of AI. This is hardly surprising given the scope for useful and commercial applications and the potential for economic gains in multiple sectors of the economy.[39] However, a few organizations have made the development of AGI their primary goal. Among the most well-known of these are DeepMind and OpenAI, each of which has demonstrated results of increasing generality, though still a long way from AGI.

What opportunities and challenges does AI pose for society?

Perhaps the most widely discussed societal impact of AI and automation is on jobs and the future of work. This is not new. In 1964, in the wake of the era's excitement about AI and automation, and concerns about their impact on jobs, President Lyndon Johnson empaneled a National Commission on Technology, Automation, and Economic Progress.[40] Among the commission's conclusions was that such technologies were important for economic growth and prosperity and "the basic fact that technology destroys jobs, but not work." Most recent studies of this effect, including those I have been involved in, have reached similar conclusions and that over time, more jobs are gained than are lost. These studies highlight that it is the sectoral and occupational transitions, the skill and wage effects – not the existence of jobs broadly – that will present the greatest challenges.[41] In their essay "Automation, AI & Work," Laura Tyson and John Zysman discuss these implications for work and workers. Michael Spence goes further, in "Automation, Augmentation, Value Creation & the Distribution of Income & Wealth," to discuss the distributional issues with respect to income and wealth within and between countries, as well as the societal opportunities that are created, especially in developing countries. In "The Turing Trap: The Promise & Peril of Human-Like Artificial Intelligence," Erik Brynjolfsson discusses how the use of human benchmarks in the development of AI runs the risk of AI that substitutes for, rather than complements, human labor. He concludes that the direction AI's development

will take in this regard, and resulting outcomes for work, will depend on the incentives for researchers, companies, and governments.[42]

Still, a concern remains that the conclusion that more jobs will be created than lost draws too much from patterns of the past and does not look far enough into the future and at what AI will be capable of. The arguments for why AI could break from past patterns of technology-driven change include: first, that unlike in the past, technological change is happening faster and labor markets (including workers) and societal systems' ability to adapt are slow and mismatched; and second, that, until now, automation has mostly mechanized physical and routine tasks, but that going forward, AI will be taking on more cognitive and nonroutine tasks, creative tasks, tasks based on tacit knowledge, and, if early examples are any indication, even socioempathic tasks are not out of the question.[43] In other words, "There are now in the world machines that think, that learn and that create. Moreover, their ability to do these things is going to increase rapidly until – in a visible future – the range of problems they can handle will be coextensive with the range to which the human mind has been applied." This was Herbert Simon and Allen Newell in 1957.[44]

Acknowledging that this time could be different usually elicits two responses: First, that new labor markets will emerge in which people will value things done by other humans for their own sake, even when machines may be capable of doing these things as well as or even better than humans. The other response is that AI will create so much wealth and material abundance, all without the need for human labor, and the scale of abundance will be sufficient to provide for everyone's needs. And when that happens, humanity will face the challenge that Keynes once framed: "For the first time since his creation man will be faced with his real, his permanent problem – how to use his freedom from pressing economic cares, how to occupy the leisure, which science and compound interest will have won for him, to live wisely and agreeably and well."[45] However, most researchers believe that we are not close to a future in which the majority of humanity will face Keynes's challenge, and that until then, there are other AI- and automation-related effects that must be addressed in the labor markets now and in the near future, such as inequality and other wage effects, education, skilling, and how humans work alongside increasingly capable machines – issues that Laura Tyson and John Zysman, Michael Spence, and Erik Brynjolfsson discuss in this volume.

Jobs are not the only aspect of the economy impacted by AI. Russell provides a directional estimate of the potentially huge economic bounty from artificial general intelligence, once fully realized: a global GDP of $750 trillion, or ten times today's global GDP. But even before we get to fully realized general-purpose AI, the commercial opportunities for companies and, for countries, the potential productivity gains and economic growth as well as economic competitiveness from narrow AI and its related technologies are more than sufficient to ensure intense

pursuit and competition by companies and countries in the development, deployment, and use of AI. At the national level, while many believe the United States is ahead, it is generally acknowledged that China is fast becoming a major player in AI, as evidenced by its growth in AI research, infrastructure, and ecosystems, as highlighted in several reports.[46] Such competition will likely have market structure effects for companies and countries, given the characteristics of such technologies as discussed by Eric Schmidt, Spence, and others elsewhere.[47] Moreover, the competitive dynamics may get in the way of responsible approaches to AI and issues requiring collective action (such as safety) between competitors, whether they are companies or countries, as Amanda Askell, Miles Brundage, and Gillian Hadfield have highlighted.[48]

Nations have reasons beyond the economic to want to lead in AI. The role of AI in national security – in surveillance, signals intelligence, cyber operations, defense systems, battle-space superiority, autonomous weapons, even disinformation and other forms of sociopolitical warfare – is increasingly clear. In "AI, Great Power Competition & National Security," Eric Schmidt, who cochaired the U.S. National Security Commission on Artificial Intelligence, paints a stark picture of current and future risks that AI technologies pose to international security and stability. Schmidt calls for the exploration of shared limits and treaties on AI, even among rivals. Short of that, he points to confidence-building measures to limit risks and increase trust.[49] At the same time, Russell and Shadbolt, respectively, spotlight concerns regarding autonomous weapons and weaponized AI.

In "The Moral Dimension of AI-Assisted Decision-Making: Some Practical Perspectives from the Front Lines," former Secretary of Defense Ash Carter identifies lessons for AI drawn from other national security-related technologies, such as nuclear weapons, while focusing on the ethics of automated decision-making. However, there are important differences between AI and nuclear technologies: for example, AI's development has been led by a private sector in pursuit of global opportunities. And, as Schmidt points out, AI technologies in their development and use have network effects and tend to consolidate around those who lead in their development, whether they are companies or countries. This pits commercial and economic interests for companies and countries on one hand, and the national security interests of countries on the other.[50] Not fully explored in this volume are the implications for companies (as well as other types of organizations) and countries not at the forefront of AI's development but that could benefit from its use. This is of particular significance given that many have highlighted the potential for AI and its related technologies to contribute, along with other social and developmental efforts, to tackling many current and future global and societal challenges.[51] The COVID-19 pandemic has given us a live example of the human cost when countries at the forefront of a globally valuable discovery, such as a vaccine, do not or are slow to share it with poorer parts of the world.

As the use of AI has grown to encompass not only consumer applications and services, but also those in health care, financial services, public services, and commerce generally, it has in many instances improved effectiveness and decision quality and enabled much-needed cost and performance optimization. At the same time, in some cases, the use of algorithms has led to issues of bias and fairness, often the result of bias in the training data and the societal systems through which such data are collected.[52] Sonia Katyal uses examples from facial recognition, policing, and sentencing to argue in "Democracy & Distrust in an Era of Artificial Intelligence" that, when there is an absence of representation and participation, AI-powered systems carry the same risks and potential for distrust as political systems. In "Distrust of Artificial Intelligence: Sources & Responses from Computer Science & Law," Cynthia Dwork and Martha Minow highlight the absence of ground truth and what happens when utility for users and commercial interests are at odds with considerations of privacy and the risks of societal harms.[53] In light of these concerns, as well as the beneficial possibilities of AI, Mariano-Florentino Cuéllar, a former California Supreme Court Justice, and Aziz Huq frame how we might achieve the title of their essay: artificially intelligent regulation.

It is easy to see how governments and organizations in their desire to observe, analyze, and optimize everything would be tempted to use AI to create increasingly powerful "seeing rooms." In "Socializing Data," Diane Coyle discusses the history and perils of seeing rooms, even when well intentioned, and the problems that arise when markets are the primary mechanism for how AI uses social data. For governments, the opportunity to use AI to improve the delivery and effectiveness of public services is also hard to ignore. In her essay "Rethinking AI for Good Governance," Helen Margetts asks what a public sector AI would look like. She draws on public sector examples from different countries to highlight key challenges, notably those related to issues like resource allocation, that are more "normatively loaded" in the public sector than they are for firms. She concludes by exploring how and in which areas governments can make the most ambitious and societally beneficial use of AI.

How much about AI is really about us?

At the end of her essay, Katyal quotes J. David Bolter from his 1984 *Dædalus* essay: "I think artificial intelligence will grow in importance as a way of looking at the human mind, regardless of the success of the programs themselves in imitating various aspects of human thought." Taking this suggestion, one can ask various kinds of questions about us using the mirror AI provides, especially as it becomes more capable: What does it mean to be intelligent, creative, or, more generally, cognitively human when many of the ways we have defined these characteristics of ourselves increasingly can be imitated or even, in the future, done better or

better done by machines? How much of being human needs the mystery of not knowing how it works, or relies on our inability to mimic it or replicate it artificially? What happens when this changes? To what extent do our human ability–bounded conceptions of *X* (where *X* could be intelligence, creativity, empathy, relations, and so on) limit the possibility of other forms of *X* that may complement or serve humanity better? To what extent must we reexamine our socioeconomic systems and institutions, our social infrastructure, what lies at the heart of our social policies, at our notions of justice, representation, and inclusion, and face up to what they really are (and have been) and what they will need to be in the age of AI?

Their shortcomings notwithstanding, the emergence of large language models and their ability to generate human-like outputs provides a "laboratory" of sorts, as Tobias Rees calls it, to explore questions about us in an era of increasingly capable machines. We may have finally arrived at what Dennett suggests at the end of his 1988 essay, that "AI has not yet solved any of our ancient riddles . . . but it has provided us with new ways of disciplining and extending philosophical imagination that we have only just begun to exploit."[54] Murati explores how humans could relate to and work alongside machines when machines can generate outputs approaching human-like creativity. She illustrates this with examples generated by GPT-3, OpenAI's large language model. The possibilities she describes echo what Scott suggests: that we humans may have to rethink our relation to work and other creative activities.

Blaise Agüera y Arcas explores the titular question of his essay "Do Large Language Models Understand Us?" through a series of provocations interspersed with outputs from LaMDA, Google's large language model. He asks whether we are gatekeeping or constantly moving the goalposts when it comes to notions such as intelligence or understanding, even consciousness, in order to retain these for ourselves. Pamela McCorduck, in her still-relevant history of the field, *Machines Who Think*, first published in 1979, put it thus: "It's part of the history of the field of artificial intelligence that every time somebody figured out how to make a computer do something – play good checkers, solve simple but relatively informal problems – there was a chorus of critics to say, 'that's not thinking.'"[55] As to what machines are actually doing or not actually doing when they appear to be thinking, one could ask whether whatever they are doing is different from what humans do in any way other than how it is being done. In "Non-Human Words: On GPT-3 as a Philosophical Laboratory," while engaging in current debates about the nature of these models, Rees also discusses how conceptions of the human have been intertwined with language in different historical eras and considers the possibility of a new era in which language is separated from humans.

In "Signs Taken for Wonders: AI, Art & the Matter of Race," Michele Elam illustrates how, throughout history, socially transformative technologies have played a formalizing and codifying role in our conceptions of what constitutes

humanity and who the "us" is. In how they are developed, used, and monetized, and by whom, she argues that technologies like AI have the effect of universalizing particular conceptions of what it is to be human and to progress, often at the exclusion of other ways of being human and of progressing and knowing, especially those associated with Black, Latinx, and Indigenous communities and with feminist, queer, disability, and decolonial perspectives; further highlighting the need for diversity among those involved in AI's development. Elsewhere, Timnit Gebru has clearly illustrated how, like other technologies with the potential to benefit society, AI can also worsen systematic discrimination of already marginalized groups.[56] In another example of AI as formalizer to ill-effect, Blaise Agüera y Arcas, Margaret Mitchell, and Alexander Todorov examine the use of machine learning to correlate physical characteristics with nonphysical traits, not unlike nineteenth- and twentieth-century physiognomy, and point out the harmful circular logic of essentialism that can result when AI is used as a detector of traits.[57]

Progress in AI not only raises the stakes on ethical issues associated with its application, it also helps bring to light issues already extant in society. Many have shown how algorithms and automated decision-making can not only perpetuate but also formalize and amplify existing societal inequalities, as well as create new inequalities.[58] In addition, the challenge to remove bias or code for fairness may also create the opportunity for society to examine in a new light what it means by "fair."[59] Here it is worth recalling Dennett being unimpressed by Putnam's indictment of AI, that "AI has utterly failed, over a quarter century, to solve problems that philosophy has utterly failed to solve over two millennia."[60] Furthermore, examining the role of algorithms and automated decision-making and the data needed to inform algorithms may shed light on what actually underlies society's goals and policies in the first place, issues that have begun to receive attention in the literature of algorithms, fairness, and social welfare.[61] In "Toward a Theory of Justice for Artificial Intelligence," Iason Gabriel, drawing on Rawls's theory of justice, explores the intersection of AI and distributive justice by considering the role that sociotechnical systems play. He examines issues including basic liberties and equality of opportunity to suggest that considerations of distributive justice may now need to grapple with the particularities of AI as a technological system and that could lead to some novel consequences.

And as AI becomes more powerful, a looming question becomes how to align AI with humans with respect to safety and control, goals and preferences, even values. The question of AI and control is as old as the field itself; Turing himself raised it, as Russell reminds us. Some researchers believe that concerns about these sorts of risks are overblown given the nature of AI, while others believe we are a long way away from existential control risks but that research must begin to consider approaches to the control issue and factor it into how we develop more powerful AI systems.[62] Russell proposes an approach to alignment and human compatibility

that capitalizes on uncertainty in goals and human preferences, and makes use of inverse reinforcement learning as a way for machines to learn human preferences. Elsewhere, Gabriel has discussed the range of possibilities as to what we mean by alignment with AI, with each possibility presenting its own complexities.[63] But in Gabriel, as in Russell, there are considerable normative challenges involved, along with complications due to the plasticity of human preferences.

In "Artificial Intelligence, Humanistic Ethics," John Tasioulas argues that designing AI that aligns with human preferences is one thing, but it does not obviate the need to determine what those human preferences should be in the first place. He challenges the tendency to default to preference utilitarianism and its maximization by AI developers, as well as by economic and governmental actors (who often use wealth maximization and GDP as proxies), which leads to market mechanisms dominating solutions at the expense of nonmarket values and mechanisms, echoing some of Coyle's concerns. Here again it seems that the mirror provided by more capable AI highlights, and with higher stakes, the unfinished (perhaps never to be finished) business of humanistic ethics, not unlike how AI may be pushing us to clarify fairness and serving notice that trolley problems are no longer just the stuff of thought experiments, since we are building autonomous systems that may have to make such choices.

Throughout the history of AI, we have asked: how good is it now? This question has been asked about every application from playing chess or Go, to knowing things, performing surgery, driving a car, writing a novel, creating art, independently making mathematical conjectures or scientific discoveries, or simply having a good bedside manner. In asking the question, it may be useful also to ask: compared to what? With an eye toward implications for society, one might compare AI with the humans best at the respective activity. There remain plenty of activities in which the "best" humans perform better than AI – as they likely will for the foreseeable future – and society is well served by these humans performing these activities. One might also compare with other samplings of humanity, such as the average person employed in or permitted to conduct that activity, or a randomly selected human. And here, as AI becomes more capable, is where the societal implications get more complicated. For example, do we raise permission standards for humans performing safety-critical activities to keep up with machine capabilities? Similarly, what determines when AI is good enough? A third comparison might be with respect to how co-extensive the range of AI capabilities become with those of humans – what Simon and Newell, as mentioned earlier, thought would eventually come to pass. How good AI systems become in this respect would likely herald the beginning of a new era for us and for society of the sort discussed previously. But perhaps the most important comparison is with respect to what we choose to use AI for and what we need AI to be capable of in order to benefit society. It would seem that in any such comparisons, along with how we

design, develop, and deploy AI, the societal implications are not foregone conclusions, but choices that are up to us.

Is all this worth it? If not, a logical response might be to stop everything, stop further development and deployment of AI, put the curses back in Pandora's box. This hardly seems realistic, given the huge economic and strategic stakes and the intense competition that has been unleashed between countries and between companies, not to mention the usefulness of AI to its users and the tantalizing beneficial possibilities, some already here, for society. My response to the question is a conditional yes.

At an AI conference a few years ago, I participated on a panel to which the host, Stuart Russell, posed a thought experiment. I forget the exact formulation, or even how I responded, but I have come to express it as follows:

> It's the year 2050, AI has turned out to be hugely beneficial to society and generally acknowledged as such. What happened?

This thought experiment aims to elicit the most worthwhile possibilities we achieved, the most beneficial opportunities we realized, the hard problems we solved, the risks we averted, the unintended consequences, misuses, and abuses we avoided, and the downsides we mitigated all in order to achieve the positive outcome in a not-too-distant future. In other words, it is a way of asking what we need to get right if AI is to be a net benefit to society.

The essays in this volume of *Dædalus* highlight many of the things we must get right. Drawing from these and other discussions, and a growing literature,[64] one can compile a long working list[65] whose items can be grouped as follows: The first group is related to the challenges of building AI powerful and capable enough to achieve the exciting beneficial possibilities for humanity, but also safe and without causing or worsening individual or group harms, and able to earn public trust, especially where societal stakes are high. A second set of challenges concerns focusing AI's development and use where it can make the greatest contributions to humanity – such as in health and the life sciences, climate change, overall well-being, and in the foundational sciences and in scientific discoveries – and to deliver net positive socioeconomic outcomes for all people. The *all* is all-important, given the likelihood that without purposeful attention to it, the characteristics of the resulting AI and its benefits could accrue to a few individuals, organizations, and countries, likely those leading in its development and use. The third group of challenges centers on the responsible development, deployment, use, and governance of AI. This is especially critical given the huge economic and geopolitical stakes and the intense competition for leadership in AI that has been unleashed between companies and between countries as a result. Not prioritizing responsible approaches to AI could lead to harmful and unsafe deployment and uses, out-

right misuses, many more unintended consequences, and destabilizing race conditions among the various competitors. A fourth set of challenges concerns us: how we co-evolve our societal systems and institutions and negotiate the complexities of how to be human in an age of increasingly powerful AI.

Readers of this volume will undoubtedly develop their own perspectives on what we collectively must get right if AI is to be a net positive for humanity. While such lists will necessarily evolve as our uses and societal experience with AI grow and as AI itself becomes more powerful, the work on them must not wait.

Returning to the question, is this worth it? My affirmative answer is conditioned on confronting and getting right these hard issues. At present, it seems that the majority of human ingenuity, effort, and financial and other resources are disproportionately focused on commercial applications and the economic potential of AI, and not enough on the other issues that are also critical for AI to be a net benefit to humanity given the stakes. We can change that.

AUTHOR'S NOTE

I am grateful to the American Academy for the opportunity to conceive this *Dædalus* volume on AI & Society and to bring together diverse perspectives on AI across a range of topics. On a theme as broad as this, there are without doubt many more topics and views that are missing; for that I take responsibility.

I would like to thank the Fellows of All Souls College, Oxford, where I have been a Visiting Fellow during the editing of this *Dædalus* volume. I would also like to thank my colleagues at the McKinsey Global Institute, the AI Index, and the 100-Year Study of AI at Stanford, as well as my fellow members on the National Academies of Sciences, Engineering, and Medicine Committee on Responsible Computing Research and Its Applications, for our many discussions as well as our work together that informed the shape of this volume. I am grateful for the conversations with the authors in this volume and with others, including Hans-Peter Brondmo, Gillian Hadfield, Demis Hassabis, Mary Kay Henry, Reid Hoffman, Eric Horvitz, Margaret Levi, Eric Salobir, Myron Scholes, Julie Su, Paul Tighe, and Ngaire Woods. I am grateful for valuable comments and suggestions on this introduction from Jack Clark, Erik Brynjolfsson, Blaise Agüera y Arcas, Julian Manyika, Sarah Manyika, Maithra Raghu, and Stuart Russell, but they should not be held responsible for any errors or opinions herein.

This volume could not have come together without the generous collaboration of the Academy's editorial team of Phyllis Bendell, Director of Publications and Managing Editor of *Dædalus*, who brought her experience as guide and editor, and enthusiasm from the very beginning to the completion of this effort, and Heather Struntz and Peter Walton, who were collaborative and expert copyeditors for all the essays in this volume.

ABOUT THE AUTHOR

James Manyika, a Fellow of the American Academy since 2019, is Chairman and Director Emeritus of the McKinsey Global Institute and Senior Partner Emeritus of McKinsey & Company, where he spent twenty-six years. He was appointed by President Obama as Vice Chair of the Global Development Council at the White House (2012–2017), and by two U.S. Commerce Secretaries to the Digital Economy Board and the National Innovation Board. He is a Distinguished Fellow of Stanford's Human-Centered AI Institute, a Distinguished Research Fellow in Ethics & AI at Oxford, and a Research Fellow of DeepMind. He is a Visiting Professor at Oxford University's Blavatnik School of Government. In early 2022, he joined Google as Senior Vice President for Technology and Society.

ENDNOTES

[1] The Turing Test was conceived by Alan Turing in 1950 as a way of testing whether a computer's responses are indistinguishable from those of a human. Though it is often discussed in popular culture as a test for artificial intelligence, many researchers do not consider it a test of artificial intelligence ; Turing himself called it "the imitation game." Alan M. Turing, "Computing Machinery and Intelligence," *Mind*, October 1950.

[2] "The Reith Lectures : Living with Artificial Intelligence," BBC, https ://www.bbc.co.uk/programmes/m001216k.

[3] Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, et al., "MuZero : Mastering Go, Chess, Shogi and Atari without Rules," DeepMind, December 23, 2020 ; and Christopher Berner, Greg Brockman, Brooke Chan, et al., "Dota 2 with Large Scale Deep Reinforcement Learning," arXiv (2019), https ://arxiv.org/abs/1912.06680.

[4] The Department of Energy's report on AI for science provides an extensive review of both the current state-of-the-art uses of AI in various branches of science as well as the grand challenges for AI in each. See Rick Stevens, Valerie Taylor, Jeff Nichols, et al., *AI for Science : Report on the Department of Energy (DOE) on Artificial Intelligence (AI) for Science* (Oak Ridge, Tenn. : U.S. Department of Energy Office of Scientific and Technical Information, 2020), https ://doi.org/ 10.2172/1604756. See also the Royal Society and the Alan Turing Institute, "The AI Revolution in Scientific Research" (London : The Royal Society, 2019).

[5] See Kathryn Tunyasuvunakool, Jonas Adler, Zachary Wu, et al., "Highly Accurate Protein Structure Prediction for the Human Proteome," *Nature* 596 (7873) (2021) ; Janet Thornton and colleagues discuss the contributions of AlphaFold to the life sciences, including its use in predicting the structure of some of the proteins associated with SARS-CoV-2, the virus that causes COVID-19. See Janet Thornton, Roman A. Laskowski, and Neera Borkakoti, "AlphaFold Heralds a Data-Driven Revolution in Biology and Medicine," *Nature Medicine* 27 (10) (2021).

[6] "AI Set to Exceed Human Brain Power," CNN, August 9, 2006, http ://edition.cnn.com/2006/TECH/science/07/24/ai.bostrom/.

[7] See Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner, "Machine Bias," ProPublica, May 23, 2016, https ://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing ; and Joy Buolamwini and Timnit Gebru, "Gender Shades : Intersectional Accuracy Disparities in Commercial Gender Classification," in *Proceedings*

of the 1st Conference on Fairness, Accountability and Transparency (New York: Association for Computing Machinery, 2018).

[8] See Hilary Putnam, "Much Ado About Not Very Much," *Dædalus* 117 (1) (Winter 1988): 279, https://www.amacad.org/sites/default/files/daedalus/downloads/Daedalus_Wi98_Artificial-Intelligence.pdf. In the same volume, see also Daniel Dennett's essay "When Philosophers Encounter Artificial Intelligence," in which he provides a robust response to Putman while also making observations about AI and philosophy that, with the benefit of hindsight, remain insightful today, even as the field has progressed.

[9] Robert K. Appiah, Jean H. Daigle, James M. Manyika, and Themuso Makhurane, "Modelling and Training of Artificial Neural Networks," *African Journal of Science and Technology Series B, Science* 6 (1) (1992).

[10] Founded by Eric Horvitz, the 100-Year Study of AI that I have been involved in publishes a report every five years; its most recent report takes stock of progress in AI as well as concerns as it is more widely deployed in society. See Michael L. Littman, Ifeoma Ajunwa, Guy Berger, et al., *Gathering Strength, Gathering Storms: The One Hundred Year Study on Artificial Intelligence (AI100) 2021 Study Panel Report* (Stanford, Calif.: Stanford University, 2021). Separately, at the AI Index, we provide an annual view of developments in AI. See Artificial Intelligence Index, Stanford University Human-Centered Artificial Intelligence, https://aiindex.stanford.edu/.

[11] B. J. Copeland, "Artificial Intelligence," Britannica, https://www.britannica.com/technology/artificial-intelligence (last edited December 14, 2021).

[12] David Poole, Alan Mackworth, and Randy Goebel, *Computational Intelligence: A Logical Approach* (New York: Oxford University Press, 1998).

[13] The goal-orientation in this second type of definition is considered by some also as limiting, hence variations such as Stuart Russell and Peter Norvig's, that focus on perceiving and acting. See Stuart Russell and Peter Norvig, *Artificial Intelligence: A Modern Approach*, 4th ed. (Hoboken, N.J.: Pearson, 2021). See also Shane Legg and Marcus Hutter, "A Collection of Definitions of Intelligence," arXiv (2007), https://arxiv.org/abs/0706.3639.

[14] See Marvin Minsky, *The Emotion Machine: Commonsense Thinking, Artificial Intelligence, and the Future of the Human Mind* (New York: Simon & Schuster, 2007).

[15] See Stephen Cave and Kanta Dihal, "Ancient Dreams of Intelligent Machines: 3,000 Years of Robots," *Nature* 559 (7715) (2018).

[16] Dennett, "When Philosophers Encounter Artificial Intelligence."

[17] John McCarthy, Marvin L. Minsky, Nathaniel Rochester, and Claude E. Shannon, "A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence," August 31, 1955, http://raysolomonoff.com/dartmouth/boxa/dart564props.pdf.

[18] Many of the pioneers of the current AI spring and their views are featured in Martin Ford, *Architects of Intelligence: The Truth about AI from the People Building It* (Birmingham, United Kingdom: Packt Publishing, 2018).

[19] Yoshua Bengio, Yann Lecun, and Geoffrey Hinton, "Deep Learning for AI," *Communications of the ACM* 64 (7) (2021). Reinforcement learning adds the notion of learning through sequential experiences that involve state transitions and making use of reinforcing rewards. See Richard S. Sutton and Andrew G. Barto, *Reinforcement Learning: An Introduction* (Cambridge, Mass.: MIT Press, 2018).

[20] Hubert L. Dreyfus and Stuart E. Dreyfus, "Making a Mind Versus Modeling the Brain: Artificial Intelligence Back at a Branchpoint," *Dædalus* 117 (1) (Winter 1988): 15–44, https://www.amacad.org/sites/default/files/daedalus/downloads/Daedalus_Wi98_Artificial-Intelligence.pdf.

[21] For a view on trends in performance versus benchmarks in various AI subfields, see Chapter 2 in Human-Centered Artificial Intelligence, *Artificial Intelligence Index Report 2022* (Stanford, Calif.: Stanford University, 2022), https://aiindex.stanford.edu/wp-content/uploads/2022/03/2022-AI-Index-Report_Master.pdf.

[22] At the time of developing this volume (2020–2021), the most well-known large language models included OpenAI's GPT-3, Google's LaMDA, Microsoft's MT-NLG, and DeepMind's Gopher. These models use transformer architectures first described in Ashish Vaswani, Noam Shazeer, Niki Parmar, et al., "Attention Is All You Need," arXiv (2017), https://arxiv.org/abs/1706.03762.

[23] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, et al., "On the Opportunities and Risks of Foundation Models," arXiv (2021), https://arxiv.org/abs/2108.07258.

[24] Ibid. See also Laura Weidinger, John Mellor, Maribeth Rauh, et al., "Ethical and Social Risks of Harm from Language Models," arXiv (2021), https://arxiv.org/abs/2112.04359. On toxicity, see Samuel Gehman, Suchin Gururangan, Maarten Sap, et al., "RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models," in *Findings of the Association for Computational Linguistics: EMNLP 2020* (Stroudsburg, Pa.: Association for Computational Linguistics, 2020), 3356–3369; and Albert Xu, Eshaan Pathak, Eric Wallace, et al., "Detoxifying Language Models Risks Marginalizing Minority Voices," arXiv (2014), https://arxiv.org/abs/2104.06390.

[25] Demis Hassabis, Dharshan Kumaran, Christopher Summerfield, and Matthew Botvinick, "Neuroscience-Inspired Artificial Intelligence," *Neuron* 95 (2) (2017); and Alexis T. Baria and Keith Cross, "The Brain Is a Computer Is a Brain: Neuroscience's Internal Debate and the Social Significance of the Computational Metaphor," arXiv (2021), https://arxiv.org/abs/2107.14042.

[26] See Littman, *Gathering Strength, Gathering Storms*.

[27] For an overview of trends in AI technical and ethics issues as well as AI regulation and policy, see Chapters 3 and 6, respectively, in Human-Centered Artificial Intelligence, *Artificial Intelligence Index Report 2022*. See also Mateusz Szczepański, Michał Choraś, Marek Pawlicki, and Aleksandra Pawlicka, "The Methods and Approaches of Explainable Artificial Intelligence," in *Computational Science – ICCS 2021*, ed. Maciej Paszynski, Dieter Kranzlmüller, Valeria V. Krzhizhanovskaya, et al. (Cham, Switzerland: Springer, 2021). See also Cynthia Dwork and Aaron Roth, "The Algorithmic Foundations of Differential Privacy," *Foundations and Trends in Theoretical Computer Science* 9 (3–4) (2014).

[28] For an overview of the types of efforts as well as three case studies (Microsoft, OpenAI, and OECD's observatory), see Jessica Cussins Newman, *Decision Points in AI Governance: Three Case Studies Explore Efforts to Operationalize AI Principles* (Berkeley: Center for Long-Term Cybersecurity, UC Berkeley, 2020).

[29] See Chapter 6, "Diversity in AI," in Human-Centered Artificial Intelligence, *Artificial Intelligence Index Report 2021* (Stanford, Calif.: Stanford University, 2021), https://aiindex.stanford.edu/wp-content/uploads/2021/11/2021-AI-Index-Report_Master.pdf. See also Sarah Myers West, Meredith Whittaker, and Kate Crawford, "Discriminating Systems: Gender, Race and Power in AI" (New York: AI Now Institute, 2019).

[30] Bengio et al., "Deep Learning for AI."

[31] Emily B. Bender and Alexander Koller, "Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (New York: Association for Computing Machinery, 2020).

[32] See Allan Dafoe, Edward Hughes, Yoram Bachrach, et al., "Open Problems in Cooperative AI," arXiv (2020), http://arxiv.org/abs/2012.08630; and Allan Dafoe, Yoram Barach, Gillian Hadfield, Eric Horvitz, et al., "Cooperative AI: Machines Must Learn to Find Common Ground," *Nature* 593 (2021).

[33] See the Department of Energy's report *AI for Science* for examples in several scientific fields. See also Alex Davies, Petar Veličković, Lars Buesing, et al., "Advancing Mathematics by Guiding Human Intuition with AI," *Nature* 600 (7887) (2021); and Anil Ananthaswamy, "AI Designs Quantum Physics Experiments Beyond What Any Human Has Conceived," *Scientific American*, July 2021.

[34] On AI's grand challenges, Raj Reddy posed probably the first list in his 1988 AAAI Presidential Address, "Foundations and Grand Challenges of Artificial Intelligence," *AI Magazine*, 1988. Ganesh Manni provides a useful history of AI grand challenges in "Artificial Intelligence's Grand Challenges: Past, Present, and Future," *AI Magazine*, Spring 2021.

[35] On the challenges and progress in causal reasoning, see Judea Pearl and Dana Mackenzie, *The Book of Why: The New Science of Cause and Effect* (New York: Basic Books, 2018).

[36] For example, see David Silver, Satinder Singh, Doina Precup, and Richard S. Sutton, "Reward is Enough," *Artificial Intelligence* 299 (4) (2021).

[37] In a recent paper, Chinese researchers describe an approach that has the potential to train models of up to 174 trillion parameters, a size that rivals the number of synapses in the brian (hence the claim of "brain-scale" models), on high performance supercomputers. See Zixuan Ma, Jiaao He, Jiezhong Qiu, et al., "BaGuaLu: Targeting Brain Scale Pretrained Models with over 37 Million Cores," March 2022, https://keg.cs.tsinghua.edu.cn/jietang/publications/PPOPP22-Ma%20et%20al.-BaGuaLu%20Targeting%20Brain%20Scale%20Pretrained%20Models%20w.pdf.

[38] See Nick Bostom, *Superintelligence: Paths, Dangers, Strategies* (Oxford: Oxford University Press, 2014); Max Tegmark, *Life 3.0: Being Human in the Age of AI* (New York: Knopf, 2017); and Martin Ford, *Rule of the Robots: How Artificial Intelligence Will Transform Everything* (New York: Basic Books, 2021).

[39] See Michael Chui, James Manyika, Mehdi Miremadi, et al., "Notes from the AI Frontier: Applications and Value of Deep Leaning" (New York: McKinsey Global Institute, 2018); and Jacques Bughin, Jeongmin Seong, James Manyika, et al., "Notes from the AI Frontier: Modeling the Impact of AI on the World Economy" (New York: McKinsey Global Institute, 2018). And for trends on adoption of AI in business and the economy as well as AI labor markets, see Chapter 4 in Human-Centered Artificial Intelligence, *Artificial Intelligence Index Report* 2022.

[40] See National Commission on Technology, Automation and Economic Progress, *Technology and the American Economy*, vol. 1 (Washington, D.C.: U.S. Government Printing Office, 1966), https://files.eric.ed.gov/fulltext/ED023803.pdf.

[41] James Manyika, Susan Lund, Michael Chui, et al., *Jobs Lost, Jobs Gained: What the Future of Work Will Mean for Jobs, Skills, and Wages* (New York: McKinsey Global Institute, 2017); Daron Acemoglu and Pascual Restrepo, "Artificial Intelligence, Automation and Work,"

NBER Working Paper 24196 (Cambridge, Mass.: National Bureau of Economic Research, 2018); David Autor, David Mindell, and Elisabeth Reynolds, *The Work of the Future: Building Better Jobs in an Age of Intelligent Machines* (Cambridge, Mass.: MIT Work of the Future, 2020); and Erik Brynjolfsson, "The Problem Is Wages, Not Jobs," in *Redesigning AI: Work, Democracy, and Justice in the Age of Automation*, ed. Daron Acemoglu (Cambridge, Mass.: MIT Press, 2021).

[42] See Daron Acemoglu and Pascual Restrepo, "The Wrong Kind of AI? Artificial Intelligence and the Future of Labor Demand," NBER Working Paper 25682 (Cambridge, Mass.: National Bureau of Economic Research, 2019); and Bryan Wilder, Eric Horvitz, and Ece Kamar, "Learning to Complement Humans," arXiv (2020), https://arxiv.org/abs/2005.00582.

[43] Susskind provides a broad survey of many of the arguments that AI has changed everything with respect to jobs. See Daniel Susskind, *A World Without Work: Technology, Automation, and How We Should Respond* (New York: Metropolitan Books, 2020).

[44] From their 1957 lecture in Herbert A. Simon and Allen Newell, "Heuristic Problem Solving: The Next Advance Operations Research," *Operations Research* 6 (1) (1958).

[45] John Maynard Keynes, "Economic Possibilities for Our Grandchildren," in *Essays in Persuasion* (New York: Harcourt Brace, 1932), 358–373.

[46] See our most recent annual AI Index report, Human-Centered Artificial Intelligence, *Artificial Intelligence Index Report* 2022. See also Daniel Castro, Michael McLaughlin, and Eline Chivot, "Who Is Winning the AI Race: China, the EU or the United States?" Center for Data Innovation, August 19, 2019; and Daitian Li, Tony W. Tong, and Yangao Xiao, "Is China Emerging as the Global Leader in AI?" *Harvard Business Review*, February 18, 2021.

[47] Tania Babina, Anastassia Fedyk, Alex Xi He, and James Hodson, "Artificial Intelligence, Firm Growth, and Product Innovation" (2021), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3651052.

[48] See Amanda Askell, Miles Brundage, and Gillian Hadfield, "The Role of Cooperation in Responsible AI Development," arXiv (2019), https://arxiv.org/abs/1907.04534.

[49] See Henry A. Kissinger, Eric Schmidt, and Daniel Huttenlocher, *The Age of AI: And Our Human Future* (Boston: Little, Brown and Company, 2021).

[50] Issues that we explored in a Council on Foreign Relations Taskforce on Innovation and National Security. See James Manyika and William H. McRaven, *Innovation and National Security: Keeping Our Edge* (New York: Council on Foreign Relations, 2019).

[51] For an assessment of the potential contributions of AI to many of the global development challenges, as well as gaps and risks, see Michael Chui, Martin Harryson, James Manyika, et al., "Notes from the AI Frontier: Applying AI for Social Good" (New York: McKinsey Global Institute, 2018). See also Ricardo Vinuesa, Hossein Azizpour, Iolanda Leita, et al., "The Role of Artificial Intelligence in Achieving the Sustainable Development Goals," *Nature Communications* 11 (1) (2022).

[52] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, et al., "Datasheets for Datasets," arXiv (2021), https://arxiv.org/abs/1803.09010.

[53] See also Shoshana Zuboff, *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power* (New York: Public Affairs, 2019).

[54] See Dennett, "When Philosophers Encounter Artificial Intelligence."

[55] Pamela McCorduck, *Machines Who Think: A Personal Inquiry into the History and Prospects of Artificial Intelligence*, 2nd ed. (Abingdon-on-Thames, United Kingdom: Routledge, 2004).

[56] Timnit Gebru, "Race and Gender," in *The Oxford Handbook of Ethics of AI*, ed. Markus D. Dubber, Frank Pasquale, and Sunit Das (Oxford: Oxford University Press, 2020).

[57] Blaise Agüera y Arcas, Margaret Mitchell, and Alexander Todorov, "Physiognomy in the Age of AI" (forthcoming).

[58] See Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner, "Machine Bias: There's Software Used across the Country to Predict Future Criminals. And It's Biased against Blacks," ProPublica, May 23, 2016; Virginia Eubanks, *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor* (New York: St. Martin's Press, 2018); and Maximilian Kasy and Rediet Abebe, "Fairness, Equality, and Power in Algorithmic Decision-Making," in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (New York: Association for Computing Machinery, 2021).

[59] See Sorelle A. Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian, "On the (im)Possibility of Fairness," arXiv (2016), https://arxiv.org/abs/1609.07236; and Arvind Narayanan, "Translation Tutorial: 21 Fairness Definitions and their Politics," in *Proceedings of the 2018 Conference on Fairness, Accountability, and Transparency* (New York: Association for Computing Machinery, 2018).

[60] Dennett, "When Philosophers Encounter Artificial Intelligence."

[61] See Sendhil Mullainathan, "Algorithmic Fairness and the Social Welfare Function," in *Proceedings of the 2018 ACM Conference on Economics and Computation* (New York: Association for Computing Machinery, 2018). See also Lily Hu and Yiling Chen, "Fair Classification and Social Welfare," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (New York: Association for Computing Machinery, 2020), 535–545; and Hoda Heidari, Claudio Ferrari, Krishna Gummadi, and Andreas Krause, "Fairness Behind a Veil of Ignorance: A Welfare Analysis for Automated Decision Making," *Advances in Neural Information Processing Systems* 31 (2018): 1265–1276.

[62] See discussion in Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies* (Oxford: Oxford University Press, 2014); and Stuart Russell, *Human Compatible: Artificial Intelligence and the Problem of Control* (New York: Viking, 2019).

[63] Iason Gabriel, "Artificial Intelligence, Values, and Alignment," *Minds and Machines* 30 (3) (2020).

[64] At an AI conference organized by the Future of Life Institute, we generated a list of priorities for robust and beneficial AI. See Stuart Russell, Daniel Dewey, and Max Tegmark, "Research Priorities for Robust and Beneficial Artificial Intelligence," *AI Magazine*, Winter 2015. See also the issues raised in Littman, *Gathering Strength, Gathering Storms*.

[65] Such a working list in response to the 2050 thought experiment can be found at "AI2050's Hard Problems Working List," https://drive.google.com/file/d/1IoSEnQSzftuW9-RikM760JSuP-Heauq_/view (accessed February 17, 2022).

"From So Simple a Beginning": Species of Artificial Intelligence

Nigel Shadbolt

Artificial intelligence has a decades-long history that exhibits alternating enthusiasm and disillusionment for the field's scientific insights, technical accomplishments, and socioeconomic impact. Recent achievements have seen renewed claims for the transformative and disruptive effects of AI. Reviewing the history and current state of the art reveals a broad repertoire of methods and techniques developed by AI researchers. In particular, modern machine learning methods have enabled a series of AI systems to achieve superhuman performance. The exponential increases in computing power, open-source software, available data, and embedded services have been crucial to this success. At the same time, there is growing unease around whether the behavior of these systems can be rendered transparent, explainable, unbiased, and accountable. One consequence of recent AI accomplishments is a renaissance of interest around the ethics of such systems. More generally, our AI systems remain singular task-achieving architectures, often termed narrow AI. I will argue that artificial general intelligence – able to range across widely differing tasks and contexts – is unlikely to be developed, or emerge, any time soon.

Artificial intelligence surrounds us, both as a topic of debate and a deployed technology. AI technologists, engineers, and scientists add to an ever-growing list of accomplishments; the fruits of their research are everywhere. Voice recognition software now goes unremarked upon on our smartphones and laptops and is ever present in digital assistants like Alexa and Siri. Our faces, fingerprints, gait, voices, and the flight of our fingers across a keypad can all be used to identify each and every one of us following the application of AI machine learning methods. AI increasingly plays a role in every sector of our economy and every aspect of our daily lives. From driving our cars to controlling our critical infrastructure, from diagnosing our illnesses to recommending content for our entertainment, AI is ubiquitous.

While pundits, politicians, and public intellectuals all weigh in on the benefits and potential harms of AI, its popular image is informed as much by Hollywood as Silicon Valley. Our cinematic representations often portray a dystopian future

© 2022 by Nigel Shadbolt
Published under a Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) license
https://doi.org/10.1162/DAED_a_01898

in which sentient machines have risen to oppress human beings. It is an old trope, one in which our technology threatens our humanity.

But it is important to look at the history and current actuality to understand what our AI future is likely to be. There are reasons to be optimistic: AI understood from a human-centered perspective augments our intelligence. It will even allow us to understand more about our own intelligence. Though, if we do not attend to AI ethics and proper regulation, it certainly has the potential to diminish us.

The title of this essay draws on the closing sentence of Charles Darwin's magisterial *On the Origin of Species*. Darwin gave us the means to understand how all of life, including self-aware, natural intelligence, has evolved. Evolution works over deep time, producing diverse species within rich and varied ecosystems. It produces complex systems whose operating and organizational principles we struggle to decipher and decode. AI has begun to populate specialist niches of the cyber-physical ecosystem, and species of narrow AI are able to master specific tasks. However, we face challenges on the same scale as cognitive neuroscientists in our quest to realize *artificial general intelligence* (AGI): systems able to reflectively range across widely differing tasks and contexts. Such systems remain the stuff of Hollywood films.

Alan Turing's famous 1950 *Mind* essay imagined a task in which a human evaluator had to determine, via a series of questions and answers between interlocutors, whether one or the other was in fact a machine.[1] He argued that the point at which this discrimination could not be reliably made would represent a watershed. The Turing Test (Turing himself called it the "imitation game") has assumed mythic status. Arguments rage as to whether it is anything like a sufficient test to determine intelligence. Years earlier, Turing had written another seminal paper in which he introduced the idea of a universal Turing machine, a formulation that showed that "it is possible to invent a single machine which can be used to compute any computable sequence."[2] The promise of this proof is the foundation upon which all modern computing devices rest.

The promise of computability also lay at the heart of the field baptized as *artificial intelligence* at the 1956 Dartmouth workshop. Computer scientist John McCarthy and his coauthors wrote in the original funding proposal: "The study 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."[3]

Much of the confidence embodied in the quote from this first era of AI lay in the formal and expressive power of logic and mathematics. Computers are grounded in Boolean logic, via transistors that implement simple logical functions: AND, NAND, OR, and NOR gates. These simple transistors give effect to functions that

allow us to build layer upon layer of more complex reasoning. Just two years after the Dartmouth conference, McCarthy produced LISP, a computer language for symbol processing that powered many early AI projects. These projects sought to decompose intelligent behavior into sets of functions that manipulated symbols. The *physical symbol system hypothesis* was the confident assertion that "a physical symbol system has the necessary and sufficient means for general intelligent action."[4] The symbols manipulated were representations of the rules and objects in tasks ranging from vision to natural language understanding, planning to game playing, theorem-proving to diagnostic reasoning.

By the 1970s, however, AI research ran into some strong headwinds. In the United States, Defense Advanced Research Projects Agency (DARPA) funding had been substantially reduced from its 1960s levels.[5] And in 1973, the United Kingdom saw the publication of the Lighthill report, in which Sir James Lighthill, Lucasian Professor of Mathematics at Cambridge University, argued that AI's "grandiose objectives" remained largely unmet, and called for a virtual halt to all AI research in Britain.[6]

It took a decade for funding levels to recover. However, by the 1980s and early 1990s, a new *domain-oriented* strand of AI – that is, knowledge-based or expert systems – was commercially successful. These systems once again demonstrated the considerable power of rule-based reasoning: systems that build proofs that establish the facts about a domain, or else attempt to establish whether a statement is true given the facts that are known or can be derived. Computers running rule-based or logic-based languages engage in cycles of forward or backward chaining to discover new facts or establish how new goals can be proved. Combined with methods of attaching certainty estimates to facts and rules, these systems found widespread deployment in sectors from medicine to aerospace, manufacturing to logistics.[7]

A new economy founded on knowledge-based systems was promised; Japanese, European, and U.S. funding agencies all invested heavily. Companies whose focus was on the software environments and hardware to support this knowledge-engineering approach flourished. Developments saw new programming ideas from AI percolate widely; the inclusion of structured representations – not just rules and logical formulas – to represent objects in a domain saw the widespread adoption of object-oriented programming methods that are pervasive today.

Unfortunately, inflated expectations and the challenges of maintaining large-scale knowledge-based systems led to another cycle of disenchantment. Funders and the market as well as some researchers in AI felt that "good old-fashioned AI (GOFAI)" approaches focused too much on a logicist interpretation of AI; what was needed was "nouvelle AI."[8] Increasing numbers of researchers argued that we needed to adopt a very different approach if we were really to understand the foundations of adaptive intelligent systems. They claimed that the best place to

look for these foundations were complex biological systems, in which animals possessed nervous systems with sensorimotor capabilities.

This was not a new claim. From the outset, many AI researchers were inspired by biological systems. The work of Norbert Wiener in cybernetics, and later Grey Walters, Walter Pitts, Warren McCulloch, and Frank Rosenblatt, used the nervous system as the base model. In 1958, Rosenblatt developed the perceptron, which was intended to model a neuron's behavior. Neurons receive multiple inputs from other connected neurons. The perceptron modeled this by receiving several input values. The connection for each input has a weight in the range of zero to one, and these values are randomly picked. The perceptron unit then sums the inputs, and if the sum exceeds a threshold value, a signal is sent to the output node; otherwise, no signal is sent. The perceptron can "learn" by adjusting the weights to approach the desired output. It implements an algorithm that classifies input into two possible categories. Inspired by the way neurons work together in the brain, the perceptron is a single-layer neural network.

In 1969, computer scientists Marvin Minsky and Seymour Papert showed that the perceptron was fundamentally limited in the functions it could compute. However, it turned out that more complex networks with connected neurons over multiple layers overcame these limitations. The mid-1980s saw the emergence of parallel distributed processing (PDP): an influential connectionist approach that was particularly good for pattern detection.[9] The PDP approach relied on the *backpropagation algorithm*, which determined how a machine should change its internal parameters and connection weights between each layer as the system was trained.

At the same time, biologically inspired robotics was taking nature as a template for design.[10] The goal was to construct complete systems with discrete behaviors and with the sensors and effectors that offloaded computational work to morphology. Simple animals, insects in particular, were favorite subjects of study. These highly successful biological systems would illustrate the methods and techniques that had worked well in real complex environments. *Animats* were all the rage: whether it was artificial crickets, modeled on their biological counterparts and who orient based on resonators, tubes through their hind legs that evolved to be a particular fraction of a wavelength of the call of a mate, or replicas of Sahara Desert ants that have an adaptation to part of their compound eyes, which are sensitive to polarized sky light, giving them directional orientation. The wisdom of bodies evolved over deep time continues to inform robotics design.

As AI approached the millennium, it comprised a broad set of methods to represent and reason about the world, from symbolic rules to knowledge represented subsymbolically in network connections. Some of these methods called for building adaptivity directly into the hardware of systems. The history of AI has constantly intertwined the discovery of new ways to reason and represent the world

with new programming languages and engineering paradigms. Computer science, in turn, has been enriched by these cycles of development.

Throughout, a fundamental contributor to AI's progress has been the increasing power of our computing substrate. Moore's law (processor capacity), Kryders's law (memory density), and Cooper's law (communication speed) all tell a story of exponential change. The accomplishments of AI and the digital revolution owe much to electrical and material engineers. The doubling of computing power, storage, and communication speeds every fifteen months has changed everything. Methods, techniques, and approaches previously intractable become possible.

As the millennium approached, increasing computing power that drove a range of AI methods and techniques allowed for impressive AI methods capable of searching huge problem spaces.

In a game in 1996, and then again in a tournament of six games in 1997, IBM's Deep Blue computer program beat Gary Kasparov, one of the very best chess players in history. How had this happened? And were the machines going to take over from us at the dawn of the new millennium? Twenty-five years ago, the ascendency of AI was announced along with the destruction of jobs and the imminent emergence of AGI.

Deep Blue was capable of evaluating one hundred million to two hundred million positions per second. Brute computing force, combined with heuristics, or rules of thumb, that suggest which part of the search tree is more interesting than another, led to uncannily capable behavior. Writing for *Time* magazine in 1996, Kasparov observed: "I had played a lot of computers but had never experienced anything like this. I could feel – I could smell – a new kind of intelligence across the table."[11] Our attribution of intelligence to the machine is a recurrent feature in our relationship with AI technology. The technology can literally unnerve us when superhuman performance is achieved. But the fundamental challenge in AI was, and remains, transferring ability in one task to another. Could all the insight generated and effort expended on Deep Blue be transferred to another task? This proved much harder.

The turn of the millennium saw another digital disruption that worked in AI's favor. The largest information asset in the history of humanity, the World Wide Web, provided a repository for vast amounts of machine-readable, open data and information. A limiting factor throughout the first half of AI's history had been a relative paucity of data. Whether for visual recognition, natural language understanding, or medical diagnosis, the data to drive learning in these domains were limited and expensive to acquire. The Web and Internet of Things (IoT) completely changed the situation. Billions of pages of text, billions of images, many of them labeled and annotated, and a flood of scientific and social data about every aspect of our lives became available as digital resources. Without these data resources, at scale, the last two decades of AI progress would have been inconceivable.

These data combined with increasingly powerful computers, search, rule-based systems, methods to learn from structured inputs, natural language understanding, and methods to compute confidence values from uncertain inputs to enable a new kind of composite AI system. In 2011, IBM announced a new age of *cognitive computing* with Watson: a system capable of beating the world's best human players not at a circumscribed board game, but at a general knowledge task.

YouTube videos of a computer competing against the best human players of the popular U.S. quiz game *Jeopardy* make for compelling viewing. In *Jeopardy*, contestants are presented with general knowledge clues in the form of answers, and they must phrase their responses in the form of questions. So, for the clue, "Wanted for general evil-ness; last seen at the tower of Barad-dur; it's a giant eye, folks. Kinda hard to miss," the correct response is "Who is Sauron?" The IBM Watson system appeared extraordinarily capable, reeling off question after question ranging over broad areas of knowledge across numerous categories.

This general intelligence could surely be transposed to other domains. Why not turn Watson into a physician? Once again, task transfer and generalization have turned out to be very difficult. While perhaps more adept at screening and triage, a physician's general problem-solving is full of task and context changes. Rather than replicating accomplished physicians, IBM's Watson Health has turned out AI assistants that can perform in routine tasks.[12]

Around the same time that Watson caught the world's attention, another AI capability was emerging, one that has delivered remarkable results. It is a continuation of the neural networks and connectionist tradition, using systems with many more hidden layers: deep neural networks (DNNs) implement highly optimized backpropagation algorithms and the principles of supervised, unsupervised, and reinforcement machine learning.

Founded in the United Kingdom in 2010 and acquired by Google in 2014, DeepMind has been a major contributor to the success of DNNs. Building on the work of researchers such as computer scientist Yann LeCun and colleagues, the company has realized a succession of brilliant task-achieving systems.[13] The promise of the DeepMind approach began to emerge with an essay showing mastery of a range of arcade games using reinforcement learning.[14]

In 2014, the AlphaGo project team was formed to test how well DNNs could compete at Go. By October 2015, a distributed version of AlphaGo beat European Go champion Fan Hui five to zero. The announcement was delayed until January 27, 2016, to coincide with the publication of the approach in *Nature*.[15] A feature of DeepMind's impact has been the follow-up of each significant achievement with peer-reviewed publications in the world's leading science journals.

A trio of DeepMind successes was released in rapid succession: AlphaGo, including AlphaGo Zero and AlphaZero; AlphaStar, DeepMind's AI program that

became ferociously good at the multiplayer strategy game StarCraft; and AlphaFold, a program that made dramatic inroads into a significant challenge for science – protein folding – helping scientists design the drugs of tomorrow.[16]

As ever, the exponents of hardware were in play. The Deep Blue machine that defeated Kasparov was one of the most powerful computers in the world, processing at 11 GigaFLOPS (eleven billion floating-point operations per second). The forty-eight tensor processing units that beat Lee Sedol, one of the world's strongest Go players, in 2016 ran at 11.5 PetaFLOPS, that is, eleven and a half thousand million million floating-point operations per second, one million times more powerful than Deep Blue.

With these types of DNN architecture, we are beginning to see AI systems augment, match, and, in some cases, outperform human experts in a whole host of tasks. Whether it is picking up underlying health conditions from retinal scans or classifying skin lesions as benign or malignant, having been trained on hundreds of thousands of images, DNNs are performing as well as the best human experts.[17] The methods behind these systems have rapidly become commercialized and commoditized. The major platforms offer cloud-based, machine learning services. They provide access to arrays of processors for training and running machine learning models. Companies invest huge amounts of capital in the development and acquisition of special hardware optimized for training and running machine learning models. Using very large data sets, they use prodigious amounts of compute power and energy to train very large neural network models. Generative Pre-trained Transformer 3 (GPT-3), a current state-of-the-art language model, trained on forty-five terabytes of data with 175 billion parameters, can be adapted to work on a wide range of tasks.[18] The model took huge amounts of cloud compute time and millions of dollars to produce. The result is a so-called foundations model, trained on broad data at scale and adaptable to a wide range of downstream tasks.[19] Such models like GPT-3 and BERT will increasingly power AI on-demand services.

AI-powered, on-demand services, such as voice, vision, and language recognition, are part of the service landscape from health to retail, finance to farming. The unreasonable effectiveness of narrow or task-specific AI has elicited familiar concerns, anxious questions about jobs and ethics, sovereign capabilities, market concentration, and our own potential redundancy as a species.

AI systems powered by machine learning methods have been used for predictive policing, suspect facial recognition, bail setting, and sentencing. But are we sure these are fair, nondiscriminatory, and proportionate? In China, AI systems are being used at scale to assign social credit. Is this supporting good citizens in a safe space or is it state surveillance? We can see the ethical issues piling up with the application of specific AI capabilities within important societal contexts (some of which are explored further in this issue of *Dædalus*).

Governments and large-tech companies, NGOs, multilateral organizations, think tanks, and universities have been busy writing their various AI ethical codes of conduct and practice. An article published in *Nature Machine Intelligence* in September 2019 presented a meta-analysis of eighty-four codes and ethical guidelines, revealing their top concerns.[20] The most prevalent of which was *transparency*, understood as efforts to increase explainability, interpretability, or other acts of communication and disclosure around AI algorithms. This undoubtedly has a great deal to do with the preponderance of DNNs. Layer upon layer of connected nodes, huge matrices of weights that somehow encode the decision-making of the trained system appear as complex black boxes.

When we are dealing with GOFAI expert systems or theorem-provers, we can see the explicit lines of reasoning; rules that can be recapitulated in natural language. If the patient has a white blood cell count of less than 2,500, then they have a low white blood cell count – such rules are applied in chains of reasoning – and if we want to know the reason for the determination of *leukopenia* (low white blood cell count) there it is, explicit and contestable.

The internals of a DNN present a challenge. There has been considerable technical work to explicate the black box. A whole subfield of AI comprises methods and techniques to understand what is going on, including efforts at feature visualization. There are striking examples in which the intermediate layers from input to output do appear to extract features that resemble the stages of processing involved, for example, in visual processing. But explainable AI remains a significant challenge.

Another top theme in the various ethical codes was that of *nonmaleficence* – a kind of do-no-ill – related to safety and security. Consider generalized adversarial networks (GANs). They comprise multiple neural networks: one, for example, classifying images and the second, its adversary, doing its best to find patterns that will have a high probability of being misclassified by the first. How can you be sure that the models you have trained are robust and cannot be subverted or indeed that the data you have trained them on have themselves not been subverted? There are methods in development to counter these attacks. But this is a race between competing methods. A product of the largely beneficial adoption of open-source principles within much of AI allows algorithms to be shared and improved as well as critiqued and compromised.

Current AI is not all about deep neural networks. AI progress has continued apace across a broad swath of approaches. Agent-based computing, which builds explicit models of competing and collaborating agents, has developed new game theoretic approaches to enable efficient and effective behavior in auctions, resource allocation, and many other applications. Agent-based computing has been used to model the pandemic and predict the impact of nonpharmacological interventions. Natural language processing methods have summarized large swaths of

scientific work that might be relevant to dealing with the pandemic. Knowledge graphs – explicit representations of biochemical and drug pathways – have been interrogated to find which drugs might be repurposed in dealing with the virus. Our current AI ecosystem has never been more varied and vibrant.

What of the future? We can be assured of continued progress in the underpinning computational fabric. The road maps available now already anticipate exponential increases in computer power, storage, and connectivity. In the United States, companies like Facebook, Amazon, and Google are increasing their investments in AI-enabled chips, as are their equivalents in China.

Data availability has been growing exponentially and, with ever more ubiquitous IoT devices, is expected to continue to do so. We may see more storage of data at the edge: that is, data that are stored locally on a plethora of distributed devices and not consolidated into the cloud. This trend will act as a forcing function on new kinds of distributed machine learning and federated problem-solving techniques. The pandemic has spawned increased amounts of data creation and replication, though estimates suggest that only 2 percent of what is created is persistently stored. The global installed storage capacity (estimated at 6.7 zettabytes in 2020) is many times smaller than the data ephemerally generated. Is this a lost opportunity? Could AI engines be uncovering more patterns and structures? And how are we to determine what data to keep?

We can be sure that the success of task-achieving architectures will continue. There are any number of image-based classification tasks to which AI methods can be applied, any number of text summarization and generation tasks to which natural language processing techniques are suited. As data become more densely connected across sectors and between individuals and organizations, there will be any number of roles for planning, recommendation, and optimization systems – lots of niches – to fill. In this sense, the future of AI will be about the continued digitization of services, products, and processes.

The current paradigm of DNNs faces significant challenges in addition to those of explainability, safety, and security already mentioned. One is the ongoing challenge of *distribution shift*. Problems arise because the data on which a network is trained come from a different distribution than the data used when tested or deployed: for example, facial recognition systems trained on a particular population and deployed in contexts with very different distributions. Distribution shift can arise because labels shift, or else the concepts involved in classification and prediction can change; whether it is the diagnostic criteria for mental illness or job titles, all are subject to considerable amounts of concept shift. Although much studied, distribution shift remains a real and ongoing challenge.

Another recurrent and recognized challenge is *transfer learning*. How can success in one task be generalized: that is, reusing or transferring information from previously learned tasks for the learning of new tasks. We already have various examples of transfer learning within AI: image-recognition systems trained on one domain transferred to another, language understanding models trained on huge data sets repurposed for other language processing tasks. But the challenge comes when the source task is not sufficiently related to the target task, or the transfer method is unable to leverage the relationship between the source and target tasks.

Notwithstanding these challenges, we will see spectacular convergences where data at scale, at new levels of precision and resolution, allow diagnosis, forecasting, and modeling across a swath of sectors. Where engineering continues its own exponential path of smaller, cheaper, more powerful, and more energy-efficient devices, we will see AI embedded into the fabric of our built environment, offering up the vision of intelligent infrastructure (II). Swarm-scale collaborations between many devices adapt to and directly modify their environments.

An approach dubbed physical AI (PAI), carrying on a tradition of biologically inspired AI, urges us to look at the underlying principles that have evolved through deep time to be intrinsic parts of biological adaption.[21] Processes resembling homeostasis, the regulation of body states aimed at maintaining conditions compatible with life, could be integrated with intelligent machines. Advocates of this approach suggest that such internal regulatory mechanisms and control will lead to a new class of machines that have intrinsic goals.[22] Mechanical engineering, computer science, biology, chemistry, and materials science will be foundational elements in this type of approach.

This gap in embodiment – in AI systems that are in themselves purposeless – remains a grand challenge for AI. Those who claim the imminent emergence of AGI should note that we remain far from understanding what constitutes our own general intelligence and associated self-awareness or consciousness. Intelligence is a polythetic concept that we use all the time and yet resists easy definitions. It is a graduated concept, we say that X is more intelligent than Y, and yet ordering ourselves on a linear scale misses the fact that we might excel in one sphere and have little or no capacity elsewhere. For most, general intelligence would seem to require language, learning, memory, and problem-solving. The importance of intuition, creativity, and reflective consciousness are seen as important attributes by many. The ability to survive in a complex world, to be embodied and possessed of perceptual and motor skills, is highlighted by others.

Patrick Winston, an AI pioneer and sometime director of MIT's Computer Science and AI Lab (CSAIL), once remarked that "there are lots of ways of being smart that aren't smart like us." On this view, the space of intelligent systems is likely large and multidimensional. Recent work on other minds invites us to consider biological entities that have a claim to many attributes of general adaptive

and intelligent behavior.[23] They are not writing literature or building cyclotrons, but the octopus displays a range of behaviors we could consider intelligent. This chimes with the nouvelle AI and Cambrian intelligence approach advocated by roboticist Rodney Brooks, an approach that builds situated robots in complex environments often exhibiting emergent behaviors.[24]

For others, consciousness is an essential feature of general intelligence. Consciousness, the hard problem in neuroscience, is itself a term that elicits very different responses. For some, it is an illusion, a kind of hallucination, a fiction we have built for ourselves. For others, it is a supervenient reality whose emergence we are far from understanding.

Whatever its basis, a key property of human consciousness is that we have *conceptual* self-awareness: we have abstract concepts for our physical and mental selves; my body, my mind, and my thought processes as well as an integrated sense of myself – me. A construct replete with emotions, experience, history, goals, and relationships. We are possessed of theories of mind to understand other entities and motivations in context, to be able to make sense of their actions and to interact with them appropriately. None of this is in our AI systems at present. This is not to say such awareness will never be present in future species of AI. Our own cognitive and neural architectures, the rich layering of systems, present an existence proof. But our AI systems are not yet in the world in any interesting sense.[25]

When discussing the prospect of artificial general intelligence, we tend to reserve a special place for our own variety – possessed of experiential self-awareness – and we seem particularly drawn to the symbolic expression of that experience in our language, teleological understanding of the world, and imagined future possibilities. We need to continue to interrogate our understanding of the concept of intelligence. For the foreseeable future, no variety of AI will have a reasonable claim to a sufficient range of attributes for us to ascribe them general intelligence. But this cannot be an in-principle embargo.

For some, this is a distraction from medium-term future concerns. Writing in the *Harvard Data Science Review*, Michael Jordan notes the need for artificial intelligence, intelligence augmentation, and intelligent infrastructure, a need that "is less about the realization of science-fiction dreams or superhuman nightmares, and more about the need for humans to understand and shape technology as it becomes ever more present and influential in their daily lives."[26]

The field of AI contains lively and intense debates about the relative contribution of particular approaches, methods, and techniques. From logic to statistical mechanics, rule-based systems to neural networks, an ever-increasing number of powerful, adaptive, and useful computational systems have been conceived, built, and deployed. We are building intelligent infrastructures suffused with adaptability, error correction, and "learning."

A range of remarkable AI-powered products and services have literally been placed in our hands through the agency of the supercomputers that are today's smartphones. These hand axes of the twenty-first century are general purpose, ubiquitous tools capable of transforming our physical and cyber worlds. The data and AI that power these systems and their successors will provide new services the early harbingers of which already exist.

Consider real-time machine translation (MT), in effect a digital realization of the Babel fish wonderfully imagined by Douglas Adams in his *Hitchhiker's Guide to the Galaxy*. This will be a world in which we speak and listen to one another, all the while remaining in our native languages. This exciting prospect comes with questions; for example, will it promote or diminish linguistic diversity? Modern statistical MT requires a lot of machine-readable text – the languages of the world are not equally represented in this regard. Is this fair or equitable?

The data and algorithms compiled into future generations of ultra-smartphones and embedded sensors will include an enormous range of diagnostic capabilities. The Babel fish will certainly be joined by a version of Star Trek's tricorder. Miniaturization will lead to device embedding and integration with our neurology and physiology. Nano probes and sensors will be on the alert for everything from cancer to dementia. Our own individual and collective biology will be available for real-time analysis and predictive maintenance. Neural links will interface with the brain to augment our senses, attention, and memory, even rendering our internal visualizations visible and inner speech audible. The associated privacy implications and challenges will be self-evident.

The real-time instrumentation of our environment will yield effective nowcasting; scientific and engineering advances via AI-augmented discovery and design will offer increased rates of innovation. Huge search spaces will be reviewed and interrogated, selected, and developed in drug and materials discovery; our artistic and cultural lives will be enriched by machine-generated content. These examples engender genuine excitement; AI empowering humankind. Sadly, weaponized AI will figure in our collective futures, too. Whether deployed to attack our cyber infrastructure or generate deepfakes, guide precision munitions or pilot drones, AI will have dangerous and lethal capabilities. Regulation and governance, ethics and law become essential adjuncts to our AI science and technology.

The "speciation" of AI, the filling of lots of niches in our cyber-physical world, is set to continue, from tasks in specific domains to support for us in all our daily tasks. The interpenetration of these tools and systems will surround and augment us. Our interactions with our AI systems will assume more texture and depth, at least from our perspective. We engineered our computational systems built on the promise of universal Turing machines. We started with the languages of logic and decision trees. We are now exploring the rich possibilities of machines driven by statistical inference, pattern-extraction, and learning from vast amounts of data.

The very recent possession of symbolic language and the discovery of mathematics and formal systems of computation have provided humans with the tools to build and explore new AI systems. This broad repertoire of approaches and methods remains essential. Our AI systems with their ability to represent and discover patterns in high dimensional data have as yet low dimensional embedding in the physical and digital worlds they inhabit. This thin tissue of grounding, of being in the world, represents the single largest challenge to realizing AGI. But the speciation of AI will continue: "from so simple a beginning endless forms most beautiful and most wonderful have been, and are being, evolved."

ABOUT THE AUTHOR

Nigel Shadbolt is Principal of Jesus College, Professorial Research Fellow in Computer Science at the University of Oxford, and Chairman and Cofounder of the Open Data Institute. He is the author of *The Digital Ape: How to Live (in Peace) with Smart Machines* (2019) and *The Spy in the Coffee Machine: The End of Privacy as We Know It* (2008), as well numerous papers on artificial intelligence, human-centered computing, and computational neuroscience.

ENDNOTES

[1] Alan M. Turing, "Computing Machinery and Intelligence," *Mind* 59 (236) (1950): 433–460, https://doi.org/10.1093/mind/LIX.236.433.

[2] Alan M. Turing, "On Computable Numbers, with an Application to the Entscheidungsproblem," *Proceedings of the London Mathematical Society* s2-42 (1) (1937): 230–265, https://doi.org/10.1112/plms/s2-42.1.230.

[3] John McCarthy, Marvin L. Minsky, Nathanial Rochester, and Claude E. Shannon, "A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence," August 1955, http://raysolomonoff.com/dartmouth/boxa/dart564props.pdf.

[4] Allen Newell and Herbert A. Simon, "Computer Science as Empirical Inquiry: Symbols and Search," *Communications of the ACM* 19 (3) (1976): 113–126, https://dl.acm.org/doi/10.1145/360018.360022.

[5] Daniel Crevier, *AI: The Tumultuous History of the Search for Artificial Intelligence* (New York: Basic Books, 1993).

[6] James Lighthill, "Artificial Intelligence: A General Survey," in *Artificial Intelligence: A Paper Symposium* (Great Britain: Science Research Council, 1973).

[7] Mark Stefik, *Introduction to Knowledge Systems* (San Francisco: Morgan Kaufmann, 1995).

[8] Rodney A. Brooks, "Intelligence without Representation," *Artificial Intelligence* 47 (1–3) (1991): 139–159.

[9] James L. McClelland and David E. Rumelhart, *Parallel Distributed Processing* (Cambridge, Mass.: MIT Press, 1986).

[10] Rolf Pfeifer and Christian Scheier, *Understanding Intelligence* (Cambridge, Mass.: MIT Press, 2001).

[11] Garry Kasparov, "The Day That I Sensed a New Kind of Intelligence," *Time* magazine, March 25, 1996.

[12] Eliza Strickland, "IBM Watson, Heal Thyself: How IBM Overpromised and Underdelivered on AI Health Care," *IEEE Spectrum* 56 (4) (2019): 24–31.

[13] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton, "Deep Learning," *Nature* 521 (7553) (2015): 436–444.

[14] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, et al., "Human-Level Control through Deep Reinforcement Learning," *Nature* 518 (7540) (2015): 529–533.

[15] David Silver, Aja Huang, Chris J. Maddison, et al., "Mastering the Game of Go with Deep Neural Networks and Tree Search," *Nature* 529 (7587) (2016): 484–489.

[16] David Silver, Julian Schrittwieser, Karen Simonyan, et al., "Mastering the Game of Go without Human Knowledge," *Nature* 550 (7676) (2017): 354–359; David Silver, Thomas Hubert, Julian Schrittwieser, et al., "A General Reinforcement Learning Algorithm that Masters Chess, Shogi, and Go through Self-Play," *Science* 362 (6419) (2018): 1140–1144; Oriol Vinyals, Igor Babuschkin, Wojciech Czarnecki, et al., "Grandmaster Level in StarCraft II Using Multi-Agent Reinforcement Learning," *Nature* 575 (7782) (2019): 350–354; and Andrew W. Senior, Richard Evans, John Jumper, et al., "Improved Protein Structure Prediction Using Potentials from Deep Learning," *Nature* 577 (7792) (2020): 706–710.

[17] Andre Esteva, Brett Kuprel, Roberto A. Novoa, et al., "Dermatologist-Level Classification of Skin Cancer with Deep Neural Networks," *Nature* 542 (7639) (2017): 115–118, https://doi.org/10.1038/nature21056.

[18] Tom B. Brown, Benjamin Mann, Nick Ryder, et al., "Language Models Are Few-Shot Learners," arXiv (2020), https://arxiv.org/abs/2005.14165.

[19] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, et al., "On the Opportunities and Risks of Foundation Models," arXiv (2021), https://arxiv.org/abs/2108.07258; and Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding," arXiv (2018), https://arxiv.org/abs/1810.04805.

[20] Anna Jobin, Marcello Ienca, and Effy Vayena, "The Global Landscape of AI Ethics Guidelines," *Nature Machine Intelligence* 1 (2019): 389–399, https://doi.org/10.1038/s42256-019-0088-2.

[21] Aslan Miriyev and Mirko Kovač, "Skills for Physical Artificial Intelligence," *Nature Machine Intelligence* 2 (2020): 658–660, https://doi.org/10.1038/s42256-020-00258-y.

[22] Kingson Man and Antonio Damasio, "Homeostasis and Soft Robotics in the Design of Feeling Machines," *Nature Machine Intelligence* 1 (2019): 446–452, https://doi.org/10.1038/s42256-019-0103-7.

[23] Peter Godfrey-Smith, *Other Minds: The Octopus and the Evolution of Intelligent Life* (London: William Collins, 2016); and Peter Godfrey-Smith, *Metazoa: Animal Minds and the Birth of Consciousness* (London: William Collins, 2020).

[24] Rodney A. Brooks, *Cambrian Intelligence: The Early History of the New AI* (Cambridge, Mass.: MIT Press, 1999).

[25] Ragnar Fjelland, "Why General Artificial Intelligence Will Not Be Realized," *Humanities and Social Sciences Communications* 7 (1) (2020): 1–9.

[26] Michael I. Jordan, "Artificial Intelligence–The Revolution Hasn't Happened Yet," *Harvard Data Science Review* 1 (1) (2019), https://doi.org/10.1162/99608f92.f06c6e61.

If We Succeed

Stuart Russell

Since its inception, AI has operated within a standard model whereby systems are designed to optimize a fixed, known objective. This model has been increasingly successful. I briefly summarize the state of the art and its likely evolution over the next decade. Substantial breakthroughs leading to general-purpose AI are much harder to predict, but they will have an enormous impact on society. At the same time, the standard model will become progressively untenable in real-world applications because of the difficulty of specifying objectives completely and correctly. I propose a new model for AI development in which the machine's uncertainty about the true objective leads to qualitatively new modes of behavior that are more robust, controllable, and deferential.

The central technical concept in AI is that of an agent: an entity that perceives and acts.[1] Cognitive faculties such as reasoning, planning, and learning are in the service of acting. The concept can be applied to humans, robots, software entities, corporations, nations, or thermostats. AI is concerned principally with designing the internals of the agent: mapping from a stream of raw perceptual data to a stream of actions. Designs for AI systems vary enormously depending on the nature of the environment in which the system will operate, the nature of the perceptual and motor connections between agent and environment, and the requirements of the task. AI seeks agent designs that exhibit "intelligence," but what does that mean?

In answering this question, AI has drawn on a much longer train of thought concerning rational behavior: what is the right thing to do? Aristotle gave one answer: "We deliberate not about ends, but about means.... [We] assume the end and consider how and by what means it is attained, and if it seems easily and best produced thereby."[2] That is, an intelligent or rational action is one that can be expected to achieve one's objectives.

This line of thinking has persisted to the present day. In the seventeenth century, theologian and philosopher Antoine Arnauld broadened Aristotle's theory to include uncertainty in a quantitative way, proposing that we should act to maximize the expected value of the outcome (that is, averaging the values of different possible outcomes weighted by their probabilities).[3] In the eighteenth century, Swiss mathematician Daniel Bernoulli refined the notion of value, moving it from an external quantity (typically money) to an internal quantity that he called utili-

© 2022 by Stuart Russell
Published under a Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) license
https://doi.org/10.1162/DAED_a_01899

ty.[4] French mathematician Pierre Rémond de Montmort noted that in games (decision situations involving two or more agents) a rational agent might have to act randomly to avoid being second-guessed.[5] And in the twentieth century, mathematician John Von Neumann and economist Oskar Morgenstern tied all these ideas together into an axiomatic framework: rational agents must satisfy certain properties such as transitivity of preferences (if you prefer A to B and B to C, you must prefer A to C), and any agent satisfying those properties can be viewed as having a utility function on states and choosing actions that maximize expected utility.[6]

As AI emerged alongside computer science in the 1940s and 1950s, researchers needed some notion of intelligence on which to build the foundations of the field. Although some early research was aimed more at emulating human cognition, the notion that won out was rationality: a machine is intelligent to the extent that its actions can be expected to achieve its objectives. In the standard model, we aim to build machines of this kind; we define the objectives and the machine does the rest. There are several different ways in which the standard model can be instantiated. For example, a problem-solving system for a deterministic environment is given a cost function and a goal criterion and finds the least-cost action sequence that leads to a goal state; a reinforcement learning system for a stochastic environment is given a reward function and a discount factor and learns a policy that maximizes the expected discounted sum of rewards. This general approach is not unique to AI. Control theorists minimize cost functions, operations researchers maximize rewards, statisticians minimize an expected loss function, and economists maximize the utility of individuals or the welfare of groups.

Within the standard model, new ideas have arisen fairly regularly since the 1950s, leading eventually to impressive real-world applications. Perhaps the oldest established area of AI is that of combinatorial search, in which algorithms consider many possible sequences of future actions or many possible configurations of complex objects. Examples include route-finding algorithms for GPS navigation, robot assembly planning, transportation scheduling, and protein design. Closely related algorithms are used in game-playing systems such as the Deep Blue chess program, which defeated world champion Garry Kasparov in 1997, and AlphaGo, which defeated world Go champion Ke Jie in 2017. In all of these algorithms, the key issue is efficient exploration to find good solutions quickly, despite the vast search spaces inherent in combinatorial problems.

Beginning around 1960, AI researchers and mathematical logicians developed ways to represent logical assertions as data structures as well as algorithms for performing logical inference with those assertions. Since that time, the technology of automated reasoning has advanced dramatically. For example, it is now routine to verify the correctness of VLSI (very large scale integration) chip designs before production and the correctness of software systems and cybersecurity

protocols before deployment in high-stakes applications. The technology of logic programming (and related methods in database systems) makes it easy to specify and check the application of complex sets of logical rules in areas such as insurance claims processing, data system maintenance, security access control, tax calculations, and government benefit distribution. Special-purpose reasoning systems designed to reason about actions can construct large-scale, provably correct plans in areas such as logistics, construction, and manufacturing. The most visible application of logic-based representation and reasoning is Google's Knowledge Graph, which, as of May 2020, holds five hundred billion facts about five billion entities and is used to answer directly more than one-third of all queries submitted to the Google search engine.[7]

In the 1980s, the AI community began to grapple with the uncertainty inherent in real-world observations and in knowledge acquired from humans or through machine learning. Although some rule-based expert systems adopted ad hoc calculi for representing and propagating uncertainty, probability theory became the dominant tool, largely due to the development of Bayesian networks by computer scientist Judea Pearl and others.[8] This led to the development of the first large-scale computational tools for probabilistic reasoning and to substantial cross-fertilization between AI and other fields that build on probability theory, including statistics, information theory, control theory, and operations research. Bayesian networks and related methods have been used for modeling, diagnosis, monitoring, and prediction of a wide range of complex systems, including jet engines, Mars rovers, ecological networks, and intensive care protocols. Causal networks, which extend Bayesian networks to model the effects of exogenous interventions, have clarified and facilitated the analysis of causal relationships in many empirical disciplines, especially in the social sciences.[9]

The development of probabilistic programming languages, or PPLs, provides a universal representation for probability models, meaning that any model representable in any formalism can be represented efficiently in a PPL.[10] Moreover, PPLs come with general-purpose inference algorithms, so that (in principle, at least) no algorithm development or mathematical derivations are needed when applying probability theory to a new domain. PPLs constitute one of the fastest-growing areas of AI and enable the rapid construction of enormously complex models. For example, the new monitoring system for the Comprehensive Nuclear-Test-Ban Treaty began life as a PPL model that took only a few minutes to write; while operating, it may dynamically construct internal representations involving hundreds of thousands of random variables.[11]

Alan Turing suggested that machine learning would be the most practical way to create AI capabilities.[12] The most common paradigm – one shared with statistical prediction methods – is supervised learning, wherein labeled examples are provided to a learning algorithm that outputs a predictive hypothesis with which to la-

bel unlabeled examples. Early developments in AI and in statistics proceeded separately, but both fields produced useful tools for learning low-dimensional models, with application to areas such as loan decisions, credit card fraud detection, and email spam filtering. For high-dimensional data such as images, deep convolutional networks have proved to be effective.[13] Deep learning has substantially advanced the state of the art in visual object recognition, speech recognition, and machine translation, three of the most important subfields of AI, as well as in protein folding, a key problem in molecular biology. Language models such as GPT-3 (Generative Pre-trained Transformer 3) – very large neural networks trained to predict the next word in a sequence – show intriguing abilities to respond to questions in a semantically meaningful way. Recent work has shown, however, that deep learning systems often fail to generalize robustly and are susceptible to spurious regularities in the training data.[14] Moreover, the amount of training data required to achieve a given level of performance is far greater than a human typically requires.

The algorithmic study of sequential decision-making under uncertainty began in economics and operations research.[15] Algorithms developed in these fields typically handle only small problems with up to one million states. In AI, the development of reinforcement learning (RL) has allowed researchers to address much larger problems satisfactorily, including backgammon with 10^{19} positions and Go with 10^{170} positions.[16] RL algorithms learn by experiencing state transitions and their associated rewards while updating a representation of the value of states (and possibly actions as well) or a direct representation of the decision policy. Applications of RL range from bidding in advertising markets to improving the ability of robots to grasp previously unseen objects.[17] As with supervised learning, applications of deep networks in RL may also be quite fragile.[18]

With modest advances in perception and dexterity, we can expect to see robots moving into a variety of unstructured environments, including roads, warehouses, agriculture, mining, and warfare. We may see progress on language understanding comparable to the progress on image understanding made over the last decade, which would enable high-impact applications such as intelligent personal assistants and high-quality intelligent tutoring systems. Search engines, rather than responding to keywords with URLs, would respond to questions with answers based on reading and, in a shallow sense, understanding almost everything the human race has ever written. And text would be augmented by satellite imagery, enabling computers to see every object (fifty centimeters or larger) on Earth every day, weather permitting.

Although this view is far from universally shared, I think it is likely that in the coming decade, the pendulum will swing away from a reliance on end-to-end deep learning and back toward systems composed from modular, semantically well-defined representations built on the mathematical foundations of logic and probability theory, with deep learning playing a crucial role in connecting to raw per-

ceptual data. (This approach underlies, for example, Waymo's industry-leading self-driving car project.) The reasons for this prediction are complex, but include 1) the performance problems with deep learning mentioned earlier; 2) the possibility that such problems may contribute to the failure of flagship projects such as self-driving cars; 3) the advantages, in terms of rigor, transparency, and modularity, of being able to analyze systems as possessing knowledge and reasoning with that knowledge; 4) the expressive limitations of circuit-based representations (including deep learning systems) for capturing general knowledge; 5) the essential role played by prior knowledge in enabling a learning system to generalize robustly from small numbers of examples; and 6) the enormous benefits of being able to improve the performance of systems by supplying knowledge rather than training data. It is important to understand that modular, semantically well-defined representations are not necessarily hand-engineered or inflexible: such representations can be learned from data, just as the entire edifice of science itself is a modular, semantically well-defined representation that has (ultimately) been learned from data.

Even in its present state, the technology of artificial intelligence raises many concerns as it transitions from research into widespread use. These concerns include potential misuses such as cybercrime, surveillance, disinformation, and political manipulation; the exacerbation of inequality and of many forms of bias in society; the creation and deployment of lethal autonomous weapons; and the usurpation of human roles in the economy and in social relationships.

These issues are addressed admirably in the other essays in this volume, many of which contribute to an important yet lamentably only recent trend: understanding potential applications of AI not only as technological problems to be solved, but also as existing in a social context. Success is to be measured not by the accuracy of the AI system's predictions and decisions, but by the real-world consequences of deploying the system. In other words, we need a theory of sociotechnical embedding for AI systems, somewhat analogous to the role that city planning plays for the artifacts produced by civil engineering and architecture. Absent such a theory, we are left with the market to sort through different systems and embeddings. For all sorts of reasons, including network effects and social externalities, this is unlikely to work.[19]

My concern here, however, is with the potential consequences of success in creating general-purpose AI: that is, systems capable of quickly learning to perform at a high level in any task environment where humans (or collections of humans) can perform well. General-purpose AI has been the long-term goal of the field since its inception. For example, Herbert Simon and Allen Newell, two pioneers of AI research, famously predicted in 1957: "There are now in the world machines that think, that learn and that create. Moreover, their

ability to do these things is going to increase rapidly until – in a visible future – the range of problems they can handle will be coextensive with the range to which the human mind has been applied."[20]

It would be an oversimplification to view progress in AI as occurring along a one-dimensional, numerical scale of "intelligence." While such a scale has some relevance for humans, AI capabilities in different branches of cognitive activity vary so markedly as to make a single scale completely inapplicable. For example, a search engine remembers very well and cannot plan at all; a chess program plans very well and cannot remember at all. For this reason, there will be no single moment at which AI "exceeds human intelligence." By the time that AI systems exhibit generality across all branches, direct comparisons to humans will be meaningless. Almost certainly, such systems would already far exceed human capabilities in many areas thanks to the massive speed, memory, and input bandwidth advantages of computers compared with humans.

That is not to imply that we are close to achieving general-purpose AI. Suggestions that we simply need to collect more data or acquire more computing power seem overly optimistic. For example, current natural-language systems process, in only a few days, thousands of times more text than any human has ever read, yet their understanding of language is brittle and often parrot-like. We need conceptual breakthroughs in a number of areas besides language understanding, including decision-making over long timescales and the cumulative use of knowledge in learning. These breakthroughs are inherently unpredictable. In a 1977 interview, John McCarthy, one of the earliest pioneers in AI, said, "What you want is 1.7 Einsteins and 0.3 of the Manhattan Project, and you want the Einsteins first. I believe it'll take five to 500 years."[21] This remains true today, although we have seen dramatic progress since 1977 in many areas. The vast majority of AI researchers now believe that general-purpose, human-level AI will arrive in this century.[22]

Given the huge levels of investment in AI research and development and the influx of talented researchers into the field, it is reasonable to suppose that fundamental advances will continue to occur as we find new applications for which existing techniques and concepts are inadequate. As noted above, these advances are hard to predict, but there are no fundamental obstacles that prevent them from occurring. Indeed, what evidence could there be that no physically possible arrangement of atoms can outperform the human brain?

The potential benefits of general-purpose AI would be far greater than those of a collection of narrow, application-specific AI systems. For this reason, the prospect of creating general-purpose AI is driving massive investments and geopolitical rivalries.

One can speculate about solving major open problems, such as extending human life indefinitely or developing faster-than-light travel, but these staples of sci-

ence fiction are not yet the driving force for progress in AI. Consider, instead, a more prosaic goal: raising the living standard of everyone on Earth, in a sustainable way, to a level that would be considered respectable in a developed country. Choosing "respectable" (somewhat arbitrarily) to mean the eighty-eighth percentile in the United States, this goal represents an almost tenfold increase in global GDP, from $76 trillion to $750 trillion per year. The increased income stream resulting from this achievement has a net present value of $13.5 quadrillion, assuming a discount factor of 5 percent. (The value is $9.4 quadrillion or $6.8 quadrillion if the technology is phased in over ten or twenty years.) These numbers tower over the amounts currently invested in AI research, and momentum toward this goal will increase as technical advances bring general-purpose AI closer to realization.

Such a tenfold increase in global GDP per capita took place over 190 years, from 1820 to 2010.[23] It required the development of factories, machine tools, automation, railways, steel, cars, airplanes, electricity, oil and gas production, telephones, radio, television, computers, the Internet, satellites, and many other revolutionary inventions. The tenfold increase in GDP posited above is predicated not on further revolutionary technologies but on the ability of general-purpose AI systems to employ what we already have more effectively and at greater scale. There would be no need to employ armies of specialists in different disciplines, organized into hierarchies of contractors and subcontractors, to carry out a project. All embodiments of general-purpose AI would have access to all the knowledge and skills of the human race, and more besides. The only differentiation would be in the physical capabilities: dexterous legged robots for construction or surgery, wheeled robots for large-scale goods transportation, quadcopter robots for aerial inspections, and so on. In principle – politics and economics aside – everyone could have at their disposal an entire organization composed of software agents and physical robots, capable of designing and building bridges or (fully automated) factories, improving crop yields, cooking dinner for one hundred guests, running elections, teaching children to read, or doing whatever else needs doing. It is the generality of general-purpose intelligence that makes this possible.

The political and economic difficulties should not, of course, be underestimated. Corporations, elites, or countries may attempt to hoard general-purpose AI technology and its benefits and, under some circumstances, economic incentives may operate to retard the dissemination of AI-based goods and services.[24] One can also expect finite resources such as land, human attention, and perhaps raw materials to become relatively more expensive.

The incentives for further development of AI, then, are huge, and the momentum appears unstoppable. We must, therefore, ask, "What if we succeed?" This question is seldom considered in the AI literature, which is focused primarily on the pursuit of success rather than on its consequences. Alan

Turing, widely regarded as the founder of computer science, did consider the question. And in 1951, during a lecture given to a learned society in Manchester, he answered: "It seems probable that once the machine thinking method had started, it would not take long to outstrip our feeble powers.... At some stage therefore we should have to expect the machines to take control."[25]

Turing's prediction is a natural response to the following conundrum: our intelligence gives us power over the world and over other species; we will build systems with superhuman intelligence; therefore, we face the problem of retaining power, forever, over entities that are far more powerful than ourselves.

Within the standard model of AI, the meaning of "power" is clear: the ability to achieve one's objectives regardless of the objectives and actions of others. I believe the future Turing had in mind was one in which machines take control as a result of pursing fixed objectives that are misaligned with human benefit. These fixed objectives will be ones that we ourselves have inserted: there is no need to posit some form of emergent consciousness that spontaneously generates its own objectives. All that is needed to assure catastrophe is a highly competent machine combined with humans who have an imperfect ability to specify human preferences completely and correctly. This is why, when a genie has granted us three wishes, our third wish is always to undo the first two wishes.

Unfortunately, the standard model within which almost all current AI systems are developed makes this future almost inevitable. Once AI systems move out of the laboratory (or artificially defined environments such as the simulated Go board) and into the real world, there is very little chance that we can specify our objectives completely and correctly in such a way that the pursuit of those objectives by more capable machines is guaranteed to result in beneficial outcomes for humans. Indeed, we may lose control altogether, as machines take preemptive steps to ensure that the stated objective is achieved.

The standard model, then, despite all its achievements, is a mistake. The mistake comes from transferring a perfectly reasonable definition of intelligence from humans to machines. It is not rational for humans to deploy machines that pursue fixed objectives when there is a significant possibility that those objectives diverge from our own.

A more sensible definition of AI would have machines pursuing *our* objectives. Of course, our objectives – in more technical language, our preferences among lotteries over complete futures – are in us, and not in the machines. This means that machines will necessarily be *uncertain* about our objectives, while being obliged to pursue them on our behalf. In this pursuit, they will be aided by evidence concerning human preferences. This evidence comes from human behavior, broadly construed, including choices, inaction, commands, requests, guidance, permissions, artifacts, and social structures.

This new model for AI, with its emphasis on uncertainty about objectives, entails a binary coupling between machines and humans that gives it a flavor quite different from the unary standard model of decoupled machines pursuing fixed objectives. The standard model can be viewed as an extreme special case of the new model, applicable only when it is reasonable to suppose that, within the machine's scope of action, the relevant human objectives can be specified completely and correctly. It turns out that the uncertainty inherent in the new model is crucial to building AI systems of arbitrary intelligence that are provably beneficial to humans.

Uncertainty concerning objectives is a surprisingly understudied topic. In the 1980s, the AI community acknowledged the inevitability of uncertainty concerning the current state and the effects of actions, but we continued to assume perfect knowledge of the objective. For artificially defined puzzles and games, this may be appropriate, but for other problems, such as recommending medical treatments, it is clear that the relevant preferences (of patients, families, doctors, insurers, hospital systems, taxpayers, and so on) are not known initially in each case. While it is true that *unresolvable* uncertainty over objectives can be integrated out of any decision problem, leaving an equivalent decision problem with a definite (average) objective, this transformation is invalid when additional evidence of the true objectives can be acquired. Thus, one may characterize the primary difference between the standard and new models of AI through the flow of preference information from humans to machines at "run-time."

This basic idea is made more precise in the framework of assistance games, originally known as cooperative inverse reinforcement learning (CIRL) games.[26] The simplest case of an assistance game involves two agents, one human and the other a robot. It is a game of partial information because, while the human knows the reward function, the robot does not, even though the robot's job is to maximize it. In a Bayesian formulation, the robot begins with a prior probability distribution over the human reward function and updates it as the robot and human interact during the game. Assistance games can be generalized to allow for imperfectly rational humans, humans who do not know their own preferences, multiple human participants, and multiple robots, among other variations.[27] Human actions in such games can, of course, include communicative actions such as stating preferences, making requests, and issuing commands.

Assistance games are connected to inverse reinforcement learning (IRL) because the robot can learn more about human preferences from the observation of human behavior – a process that is the dual of reinforcement learning, wherein behavior is learned from rewards and punishments.[28] The primary difference is that in the assistance game, unlike the IRL framework, the human's actions are affected by the robot's presence. For example, the human may try to teach the robot about their preferences, and the robot may interpret the human's actions in this light, rather than simply as demonstrations of optimal behavior.

Within the framework of assistance games, a number of basic results can be established that are relevant to Turing's problem of control.

- Under certain assumptions about the support and bias of the robot's prior probability distribution over human rewards, one can show that a robot solving an assistance game has nonnegative value to humans.[29]
- A robot that is uncertain about the human's preferences has a nonnegative incentive to allow itself to be switched off.[30] In general, it will defer to human control actions.
- To avoid changing attributes of the world whose value is unknown, the robot will generally engage in "minimally invasive" behavior to benefit the human.[31] Even when it knows nothing at all about human preferences, it will still take "empowering" actions that expand the set of actions available to the human.

Needless to say, there are many open research problems in the new model of AI. First, we need to examine each existing research area (search, game playing, constraint satisfaction, planning, reinforcement learning, and so on) and remove the assumption of a fixed, known objective, rebuilding that area on a broader foundation that allows for uncertainty about objectives. The key questions in each area are how to formulate the machine's initial uncertainty about human preferences and how to codify the run-time flow of preference information from human to machine.

Another set of research problems arises when we consider how the machine can learn about human preferences from human behavior in the assistance game. The first difficulty is that humans are irrational in the sense that our actions do not reflect our preferences. This irrationality arises in part from our computational limitations relative to the complexity of the decisions we face. For example, if two humans are playing chess and one of them loses, it is because the loser (and possibly the winner, too) made a mistake, a move that led inevitably to a forced loss. A machine observing that move and assuming perfect rationality on the part of the human might well conclude that the human *preferred* to lose. Thus, to avoid reaching such conclusions, the machine must take into account the *actual* cognitive mechanisms of humans.

Another important consequence of human computational limitations is that they force us to organize our behavior hierarchically. That is, we make (defeasible) commitments to higher-level goals such as "write an essay on a human-compatible approach to AI." Then, rather than considering all possible sequences of words, from "aardvark aardvark aardvark" to "zyzzyva zyzzyva zyzzyva," as a chess program might do, we choose among subtasks such as "write the introduc-

tion" and "read more about preference elicitation." Eventually, we get down to the choice of words, and then typing each word involves a sequence of keystrokes, each of which is in turn a sequence of motor control commands to the muscles of the arms and hands. At any given point, then, a human is embedded at various particular levels of multiple deep and complex hierarchies of partially overlapping activities and subgoals. This means that for the machine to understand human actions, it probably needs to understand a good deal about what these hierarchies are and how we use them to navigate the real world.

Other research problems engage directly with philosophy and the social sciences. For example, there is the question of social aggregation, a staple of economics and moral philosophy: how should a machine make decisions when its actions affect the interests of more than one human being? Issues include the preferences of evil individuals, relative preferences and positional goods, and interpersonal comparison of preferences.[32]

Also of great importance is the plasticity of human preferences: the fact that they seem to change over time as the result of experiences. It is hard to explain how such changes can be made rationally, since they make one's future self less likely to satisfy one's present preferences about the future. Yet plasticity seems fundamentally important to the entire enterprise, because newborn infants certainly lack the rich, nuanced, culturally informed preference structures of adults. Indeed, it seems likely that our preferences are at least partially formed by a process resembling inverse reinforcement learning, whereby we absorb preferences that explain the behavior of those around us. Such a process would tend to give cultures some degree of autonomy from the otherwise homogenizing effects of our dopamine-based reward system.

Plasticity also raises the obvious question of which human H the machine should try to help: H_{2022}, H_{2035}, or some time-averaged H?[33] Plasticity is also problematic because of the possibility that the machine may, by subtly influencing the environment, gradually mold H's preferences in directions that make them easier to satisfy. This problem is a familiar one in human society, where culture and propaganda mold the preferences of humans to facilitate their compliance with existing power structures.

Let us assume, for the sake of argument, that all these obstacles can be overcome, as well as all of the obstacles to the development of truly capable AI systems. Are we then home free? Would provably beneficial, superintelligent AI usher in a golden age for humanity? Not necessarily. There remains the issue of adoption: how can we obtain broad agreement on suitable design principles, and how can we ensure that only suitably designed AI systems are deployed?

On the question of obtaining agreement at the policy level, it is necessary first to generate consensus within the research community on the basic ideas of – and

design templates for – provably beneficial AI, so that policy-makers have some concrete guidance on what sorts of regulations might make sense. Economic incentives would tend to support the installation of rigorous standards at the early stages of AI development, since failures would be damaging to entire industries, not just to the perpetrator and victim. We already see this in miniature with the imposition of machine-checkable software standards for cell phone applications.

On the question of enforcement, I am less sanguine. If the next Dr. Evil wants to take over the world, he or she might remove the safety catch, so to speak, and deploy a poorly designed AI system that ends up destroying the world instead. This is a hugely magnified version of the problem we currently face with malware. Our track record in solving the latter problem does not provide grounds for optimism concerning the former. In Samuel Butler's *Erewhon* and in Frank Herbert's *Dune*, the solution is to ban all intelligent machines, as a matter of both law and cultural imperative. Perhaps if we find institutional solutions to the malware problem, we will be able to devise some less drastic approach for regulating AI.

The problem of misuse is not limited to evil masterminds. One possible future for humanity in the age of superintelligent AI is that of a race of lotus eaters, progressively enfeebled as machines take over the management of our entire civilization. This is the future imagined in E. M. Forster's story *The Machine Stops*, written in 1909. We may say, now, that such a future is undesirable; the machines may agree with us and volunteer to stand back, requiring humanity to exert itself and maintain its vigor. But exertion is tiring, and we may, in our usual myopic way, design AI systems that are not quite so concerned about the long-term vigor of humanity and are just a little more helpful than they would otherwise wish to be. Unfortunately, this slope is very slippery indeed.

Finding a solution to the AI control problem is an important task; it may be, in the words of philosopher Nick Bostrom, "the essential task of our age."[34] Up to now, AI research has focused on systems that are better at making decisions, but this is not the same as making better decisions if human and machine objectives diverge.

This problem requires a change in the definition of AI itself: from a field concerned with a unary notion of intelligence as the optimization of a given objective to a field concerned with a binary notion of machines that are provably beneficial for humans. Taking the problem seriously seems likely to yield new ways of thinking about AI, its purpose, and our relationship with it.

ABOUT THE AUTHOR

Stuart Russell is Professor of Computer Science and the Smith-Zadeh Professor in Engineering at the University of California, Berkeley, and Honorary Fellow at Wadham College, Oxford. He is the author of *Artificial Intelligence: A Modern Approach* (with Peter Norvig, 4th ed., 2021), *Human Compatible: AI and the Problem of Control* (2019), and *Do the Right Thing: Studies in Limited Rationality* (with Eric H. Wefald, 1991).

ENDNOTES

[1] Stuart Russell and Peter Norvig, *Artificial Intelligence: A Modern Approach*, 4th ed. (Hoboken, N.J.: Pearson, 2021).

[2] Aristotle, *Nicomachean Ethics* 3.3, 1112b.

[3] Antoine Arnauld, *La logique, ou l'art de penser* (Paris: Chez Charles Savreux, 1662).

[4] Daniel Bernoulli, "Specimen theoriae novae de mensura sortis," *Proceedings of the St. Petersburg Imperial Academy of Sciences* 5 (1738): 175–192.

[5] Pierre Rémond de Montmort, *Essay d'analyse sur les jeux de hazard*, 2nd ed. (Paris: Chez Jacques Quillau, 1713).

[6] John Von Neumann and Oskar Morgenstern, *Theory of Games and Economic Behavior* (Princeton, N.J.: Princeton University Press, 1944).

[7] Danny Sullivan "A Reintroduction to Our Knowledge Graph and Knowledge Panels," The Keyword, Google, May 20, 2020, https://blog.google/products/search/about-knowledge-graph-and-knowledge-panels/.

[8] Judea Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference* (Burlington, Mass.: Morgan Kaufmann, 1988).

[9] Judea Pearl, *Causality: Models, Reasoning, and Inference* (Cambridge: Cambridge University Press, 2000); and Judea Pearl and Dana Mackenzie, *The Book of Why: The New Science of Cause and Effect* (New York: Basic Books, 2018).

[10] Daphne Koller, David McAllester, and Avi Pfeffer, "Effective Bayesian Inference for Stochastic Programs," in *Proceedings of Fourteenth National Conference on Artificial Intelligence* (Menlo Park, Calif.: Association for the Advancement of Artificial Intelligence, 1997); Avi Pfeffer, "IBAL: A Probabilistic Rational Programming Language," in *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence* (Santa Clara, Calif.: International Joint Conference on Artificial Intelligence Organization, 2001); Brian Milch, Bhaskara Marthi, Stuart Russell, et al., "BLOG: Probabilistic Models with Unknown Objects," in *Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence* (Santa Clara, Calif.: International Joint Conference on Artificial Intelligence Organization, 2005); and Noah D. Goodman, Vikash K. Mansinghka, Daniel Roy, et al., "Church: A Language for Generative Models," in *Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence* (Helsinki: Association for Uncertainty in Artificial Intelligence, 2008).

[11] Ronan Le Bras, Nimar Arora, Noriyuki Kushida, et al., "NET-VISA from Cradle to Adulthood: A Machine-Learning Tool for Seismo-Acoustic Automatic Association," *Pure and Applied Geophysics* 178 (2021): 2437–2458.

[12] Alan Turing, "Computing Machinery and Intelligence," *Mind* 56 (236) (1950): 43–60.

[13] Yann LeCun, Lawrence Jackel, Bernhard Boser, et al., "Handwritten Digit Recognition: Applications of Neural Network Chips and Automatic Learning," *IEEE Communications Magazine* 27 (11) (1989): 41–46; Yann LeCun, Yoshua Bengio, and Geoffrey Hinton, "Deep Learning," *Nature* 521 (7553) (2015): 436–444; Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *Advances in Neural Information Processing Systems* 25 (2) (2012): 1097–1105.

[14] Brandon Carter, Siddhartha Jain, Jonas Mueller, and David Gifford, "Overinterpretation Reveals Image Classification Model Pathologies," arXiv (2020), https://arxiv.org/abs/2003.08907; and Alexander D'Amour, Katherine Heller, Dan Moldovan, et al., "Underspecification Presents Challenges for Credibility in Modern Machine Learning," arXiv (2020), https://arxiv.org/abs/2011.03395.

[15] Lloyd S. Shapley, "Stochastic Games," *Proceedings of the National Academy of Sciences* 39 (10) (1953): 1095–1100; Richard Bellman, "On the Theory of Dynamic Programming," *Proceedings of the National Academy of Sciences* 38 (8) (1952): 716–719; and Richard Bellman, *Dynamic Programming* (Princeton, N.J.: Princeton University Press, 1957).

[16] Arthur L. Samuel, "Some Studies in Machine Learning Using the Game of Checkers," *IBM Journal of Research and Development* 3 (3) (1959): 210–229; and David Silver, Aja Huang, Chris J. Maddison, et al., "Mastering the Game of Go with Deep Neural Networks and Tree Search," *Nature* 529 (7587) (2016): 484–489.

[17] Junqi Jin, Chengru Song, Han Li, et al., "Real-Time Bidding with Multi-Agent Reinforcement Learning in Display Advertising," in *Proceedings of the 27th ACM International Conference on Information and Knowledge Management* (New York: Association for Computing Machinery, 2018), 2193–2201; and Deirdre Quillen, Eric Jang, Ofir Nachum, et al., "Deep Reinforcement Learning for Vision-Based Robotic Grasping: A Simulated Comparative Evaluation of Off-Policy Methods," in *Proceedings of the 2018 IEEE International Conference on Robotics and Automation (ICRA)* (Piscataway, N.J.: Institute of Electrical and Electronics Engineers, 2018), 6284–6291.

[18] Adam Gleave, Michael Dennis, Neel Kant, et al., "Adversarial Policies: Attacking Deep Reinforcement Learning," in *Proceedings of the Eighth International Conference on Learning Representations* (La Jolla, Calif.: International Conference on Learning Representations, 2020).

[19] Eric Posner and Glen Weyl, *Radical Markets: Uprooting Capitalism and Democracy for a Just Society* (Princeton, N.J.: Princeton University Press, 2019).

[20] Herbert A. Simon and Allen Newell, "Heuristic Problem Solving: The Next Advance in Operations Research," *Operations Research* 6 (1) (1958).

[21] Israel Shenker, "Brainy Robots in Our Future, Experts Think," *Detroit Free Press*, September 30, 1977.

[22] Katja Grace, John Salvatier, Allan Dafoe, et al., "When Will AI Exceed Human Performance? Evidence from AI Experts," *Journal of Artificial Intelligence Research* (62) (2018): 729–754.

[23] Jan Luiten Van Zanden, Joerg Baten, Marco Mira d'Ercole, et al., eds., *How Was Life? Global Well-Being Since 1820* (Paris: OECD Publishing, 2014).

[24] Philippe Aghion, Benjamin F. Jones, and Charles I. Jones, "Artificial Intelligence and Economic Growth," National Bureau of Economic Research Working Paper 23928 (Cambridge, Mass.: National Bureau of Economic Research, 2017).

[25] Alan Turing, "'Intelligent Machinery, A Heretical Theory,' a Lecture Given to '51 Society' at Manchester," AMT/B/4, The Turing Digital Archive, https://turingarchive.kings.cam.ac.uk/publications-lectures-and-talks-amtb/amt-b-4.

[26] Dylan Hadfield-Menell, Anca Dragan, Pieter Abbeel, and Stuart Russell, "Cooperative Inverse Reinforcement Learning," *Advances in Neural Information Processing Systems* 29 (2016): 3909–3917.

[27] For examples, see Dylan Hadfield-Menell, Anca Dragan, Pieter Abbeel, and Stuart Russell, "The Off-Switch Game," in *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence* (Santa Clara, Calif.: International Joint Conference on Artificial Intelligence Organization, 2017), 220–227; Lawrence Chan, Dylan Hadfield-Menell, Siddhartha Srinivasa, and Anca Dragan, "The Assistive Multi-Armed Bandit," in *14th ACM/IEEE International Conference on Human-Robot Interaction (HRI 2019)* (Red Hook, N.Y.: Curran Associates, Inc., 2019), 354–363; and Arnaud Fickinger, Simon Zhuang, Andrew Critch, et al., "Multi-Principal Assistance Games: Definition and Collegial Mechanisms," presented at the Cooperative AI Research Workshop at the 34th Conference on Neural Information Processing Systems (NeurIPS 2020), virtual conference, December 6–12, 2020.

[28] Stuart Russell, "Learning Agents for Uncertain Environments," in *Proceedings of the Eleventh ACM Conference on Computational Learning Theory* (New York: Association for Computing Machinery, 1998); and Andrew Ng and Stuart Russell, "Algorithms for Inverse Reinforcement Learning," in *Proceedings of Seventeenth International Conference on Machine Learning* (San Francisco: Morgan Kaufmann Publishers, Inc., 2000).

[29] Hadfield-Menell et al., "Cooperative Inverse Reinforcement Learning."

[30] Hadfield-Menell et al., "The Off-Switch Game."

[31] Rohin Shah, Dmitrii Krasheninnikov, Jordan Alexander, et al., "Preferences Implicit in the State of the World," in *Proceedings of the Seventh International Conference on Learning Representations* (La Jolla, Calif.: International Conference on Learning Representations, 2019).

[32] John C. Harsanyi, "Morality and the Theory of Rational Behavior," *Social Research* 44 (4) (1977): 623–656; Thorstein Veblen, *The Theory of the Leisure Class: An Economic Study of Institutions* (London: Macmillan Company, 1899); Fred Hirsch, *Social Limits of Growth* (London: Routledge and Kegan Paul, 1977); Robert Nozick, *Anarchy, State, and Utopia* (New York: Basic Books, 1974); and Amartya Sen, "The Possibility of Social Choice," *American Economic Review* 89 (3) (1999): 349–378.

[33] Richard Pettigrew, *Choosing for Changing Selves* (Oxford: Oxford University Press, 2020).

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

A Golden Decade of Deep Learning: Computing Systems & Applications

Jeffrey Dean

The past decade has seen tremendous progress in the field of artificial intelligence thanks to the resurgence of neural networks through deep learning. This has helped improve the ability for computers to see, hear, and understand the world around them, leading to dramatic advances in the application of AI to many fields of science and other areas of human endeavor. In this essay, I examine the reasons for this progress, including the confluence of progress in computing hardware designed to accelerate machine learning and the emergence of open-source software frameworks to dramatically expand the set of people who can use machine learning effectively. I also present a broad overview of some of the areas in which machine learning has been applied over the past decade. Finally, I sketch out some likely directions from which further progress in artificial intelligence will come.

Since the very earliest days of computing, humans have dreamed of being able to create "thinking machines." The field of artificial intelligence was founded in a workshop organized by John McCarthy in 1956 at Dartmouth College, with a group of mathematicians and scientists getting together to "find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves."[1] The workshop participants were optimistic that a few months of focused effort would make real progress on these problems.

The few-month timeline proved overly optimistic. Over the next fifty years, a variety of approaches to creating AI systems came into and fell out of fashion, including logic-based systems, rule-based expert systems, and neural networks.[2] Approaches that involved encoding logical rules about the world and using those rules proved ineffective. Hand-curation of millions of pieces of human knowledge into machine-readable form, with the Cyc project as the most prominent example, proved to be a very labor-intensive undertaking that did not make significant headway on enabling machines to learn on their own.[3] Artificial neural networks, which draw inspiration from real biological neural networks, seemed like a promising approach for much of this time, but ultimately fell out of favor in the 1990s. While they were able to produce impressive results for toy-scale problems, they

© 2022 by Jeffrey Dean
Published under a Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) license
https://doi.org/10.1162/DAED_a_01900

were unable to produce interesting results on real-world problems at that time. As an undergraduate student in 1990, I was fascinated by neural networks and felt that they seemed like the right abstraction for creating intelligent machines and was convinced that we simply needed more computational power to enable larger neural networks to tackle larger, more interesting problems. I did an undergraduate thesis on parallel training of neural networks, convinced that if we could use sixty-four processors instead of one to train a single neural network then neural networks could solve more interesting tasks.[4] As it turned out, though, relative to the computers in 1990, we needed about one million times more computational power, not sixty-four times, for neural networks to start making impressive headway on challenging problems! Starting in about 2008, though, thanks to Moore's law, we started to have computers this powerful, and neural networks started their resurgence and rise into prominence as the most promising way to create computers that can see, hear, understand, and learn (along with a rebranding of this approach as "deep learning").

The decade from around 2011 to the time of writing (2021) has shown remarkable progress in the goals set out in that 1956 Dartmouth workshop, and machine learning (ML) and AI are now making sweeping advances across many fields of endeavor, creating opportunities for new kinds of computing experiences and interactions, and dramatically expanding the set of problems that can be solved in the world. This essay focuses on three things: the computing hardware and software systems that have enabled this progress; a sampling of some of the exciting applications of machine learning from the past decade; and a glimpse at how we might create even more powerful machine learning systems, to truly fulfill the goals of creating intelligent machines.

Hardware and software for artificial intelligence. Unlike general-purpose computer code, such as the software you might use every day when you run a word processor or web browser, deep learning algorithms are generally built out of different ways of composing a small number of linear algebra operations: matrix multiplications, vector dot products, and similar operations. Because of this restricted vocabulary of operations, it is possible to build computers or accelerator chips that are tailored to support just these kinds of computations. This specialization enables new efficiencies and design choices relative to general-purpose central processing units (CPUs), which must run a much wider variety of kinds of algorithms.

During the early 2000s, a handful of researchers started to investigate the use of graphics processing units (GPUs) for implementing deep learning algorithms. Although originally designed for rendering graphics, researchers discovered that these devices are also well suited for deep learning algorithms because they have relatively high floating-point computation rates compared with CPUs. In 2004,

computer scientists Kyoung-Su Oh and Keechul Jung showed a nearly twenty-fold improvement for a neural network algorithm using a GPU.[5] In 2008, computer scientist Rajat Raina and colleagues demonstrated speedups of as much as 72.6 times from using a GPU versus the best CPU-based implementation for some unsupervised learning algorithms.[6]

These early achievements continued to build, as neural networks trained on GPUs outperformed other methods in a wide variety of computer vision contests.[7] As deep learning methods began showing dramatic improvements in image recognition, speech recognition, and language understanding, and as more computationally intensive models (trained on larger data sets) continued demonstrating improved results, the field of machine learning really took off.[8] Computer systems designers started to look at ways to scale deep learning models to even more computationally intensive heights. One early approach used large-scale distributed systems to train a single deep learning model. Google researchers developed the DistBelief framework, a software system that enabled using large-scale distributed systems for training a single neural network.[9] Using DistBelief, researchers were able to train a single unsupervised neural network model that was two orders of magnitude larger than previous neural networks. The model was trained on a large collection of random frames from YouTube videos, and with a large network and sufficient computation and training data, it demonstrated that individual artificial neurons (the building blocks of neural networks) in the model would learn to recognize high-level concepts like human faces or cats, despite never being given any information about these concepts other than the pixels of raw images.[10]

These successes led system designers to design computational devices that were even better suited and matched to the needs of deep learning algorithms than GPUs. For the purpose of building specialized hardware, deep learning algorithms have two very nice properties. First, they are very tolerant of reduced precision. Unlike many numerical algorithms, which require 32-bit or 64-bit floating-point representations for the numerical stability of the computations, deep learning algorithms are generally fine with 16-bit floating-point representations during training (the process by which neural networks learn from observations), and 8-bit and even 4-bit integer fixed-point representations during inference (the process by which neural networks generate predictions or other outputs from inputs). The use of reduced precision enables more multiplication circuits to be put into the same chip area than if higher-precision multipliers were used, meaning chips can perform more computations per second. Second, the computations needed for deep learning algorithms are almost entirely composed of different sequences of linear algebra operations on dense matrices or vectors, such as matrix multiplications or vector dot products. This led to the observation that making chips and systems that were specialized for low-precision linear algebra computations could give very large benefits in terms of better performance per dollar and better per-

formance per watt. An early chip in this vein was Google's first Tensor Processing Unit (TPUv1), which targeted 8-bit integer computations for deep learning inference and demonstrated one to two order-of-magnitude improvements in speed and performance per watt over contemporary CPUs and GPUs.[11] Deployments of these chips enabled Google to make dramatic improvements in speech recognition accuracy, language translation, and image classification systems. Later TPU systems are composed of custom chips as well as larger-scale systems connecting many of these chips together via high-speed custom networking into pods, large-scale supercomputers designed for training deep learning models.[12] GPU manufacturers like NVIDIA started tailoring later designs toward lower-precision deep learning computations and an explosion of venture capital–funded startups sprung up building various kinds of deep learning accelerator chips, with GraphCore, Cerebras, SambaNova, and Nervana being some of the most well-known.

Alongside the rise of GPUs and other ML-oriented hardware, researchers developed open-source software frameworks that made it easy to express deep learning models and computations. These software frameworks are still critical enablers. Today, open-source frameworks help a broad set of researchers, engineers, and others push forward deep learning research and apply deep learning to an incredibly wide range of problem domains (many of which are discussed below). Some of the earliest frameworks like Torch, developed starting in 2003, drew inspiration from earlier mathematical tools like MatLab and NumPy.[13] Theano, developed in 2010, was an early deep learning–oriented framework that included automatic symbolic differentiation.[14] Automatic differentiation is a useful tool that greatly eases the expression of many gradient-based machine learning algorithms, such as stochastic gradient descent (a technique in which errors in outputs are corrected by comparing the actual output and the desired output and making small adjustments to the model parameters in the direction of the error gradient). DistBelief and Caffe were frameworks developed in the early 2010s that emphasized scale and performance.[15]

TensorFlow is a framework that allows the expression of machine learning computations.[16] It was developed and open-sourced by Google in 2015 and combines ideas from earlier frameworks like Theano and DistBelief.[17] TensorFlow was designed to target a wide variety of systems and allows ML computations to run on desktop computers, mobile phones, large-scale distributed environments in data centers, and web browsers, and targets a wide variety of computation devices, including CPUs, GPUs, and TPUs. The system has been downloaded more than fifty million times and is one of the most popular open-source packages in the world. It has enabled a tremendous range of uses of machine learning by individuals and organizations large and small all around the world.

PyTorch, released in 2016, has gained popularity with researchers for its easy expression of a variety of research ideas using Python.[18] JAX, released in 2018, is a

popular open-source Python-oriented library combining sophisticated automatic differentiation and an underlying XLA compiler, also used by TensorFlow to efficiently map machine learning computations onto a variety of different types of hardware.[19]

The importance of open-source machine learning libraries and tools like TensorFlow and PyTorch cannot be overstated. They allow researchers to quickly try ideas and express them on top of these frameworks. As researchers and engineers around the world build on each other's work more easily, the rate of progress in the whole field accelerates!

Research explosion. As a result of research advances, the growing computational capabilities of ML-oriented hardware like GPUs and TPUs, and the widespread adoption of open-source machine learning tools like TensorFlow and PyTorch, there has been a dramatic surge in research output in the field of machine learning and its applications. One strong indicator is the number of papers posted to the machine learning–related categories of arXiv, a popular paper preprint hosting service, with more than thirty-two times as many paper preprints posted in 2018 as in 2009 (a growth rate of more than double every two years).[20] There are now more than one hundred research papers posted to arXiv per day in the machine learning–related subtopic areas, and this growth shows no signs of slowing down.

Application explosion. The transformative growth in computing power, advances in software and hardware systems for machine learning, and the surge of machine learning research have all led to a proliferation of machine learning applications across many areas of science and engineering. By collaborating with experts in critical fields like climate science and health care, machine learning researchers are helping to solve important problems that can be socially beneficial and advance humanity. We truly live in exciting times.

Neuroscience is one important area in which machine learning has accelerated scientific progress. In 2020, researchers studied a fly brain to understand more about how the human brain works. They built a connectome, a synapse-resolution-level map of connectivity of an entire fly brain.[21] But without machine learning and the computational power we now have, this would have taken many years. For example, in the 1970s, it took researchers about ten years to painstakingly map some three hundred neurons within the brain of a worm. By contrast, a fly brain has one hundred thousand neurons, and a mouse brain (the next goal for machine learning–aided connectomics) has about seventy million neurons. A human brain contains about eighty-five billion neurons, with about one thousand connections per neuron. Fortunately, deep learning–based advances in computer vision now make it possible to speed up this previously gargantuan process. And

today, thanks to machine learning, you can explore the fly brain for yourself using an interactive 3-D model![22]

Molecular biology. Machine learning can also help us understand more about our genetic makeup and, ultimately, address gene-based disease more effectively. These new techniques allow scientists to explore the landscape of potential experiments much more quickly through more accurate simulation, estimation, and data analysis. One open-source tool, DeepVariant, can more accurately process the raw information coming from DNA sequencing machines (which contain errors introduced by the physical process of reading the genetic sequence) and analyze it to more accurately identify the true genetic variants in the sequence relative to a reference genome data using a convolutional neural network. Once genetic variants have been identified, deep learning can also help to analyze genetic sequences to better understand genetic features of single or multiple DNA mutations that cause particular health or other outcomes. For example, a study led by the Dana-Farber Cancer Institute improved diagnostic yield by 14 percent for genetic variants that lead to prostate cancer and melanoma in a cohort of 2,367 cancer patients.[23]

Health care. Machine learning is also offering new ways to help detect and diagnose disease. For example, when applied to medical images, computer vision can help doctors diagnose a number of serious diseases more quickly and accurately than doctors can on their own.

One impressive example is the ability for deep neural networks to correctly diagnose diabetic retinopathy, generally on par with human ophthalmologists. This ocular disease is the fastest growing cause of preventable blindness (projected to impact 642 million people in 2040).

Deep learning systems can also help detect lung cancer as well or better than trained radiologists. The same goes for breast cancer, skin disease, and other diseases.[24] The application of sequential prediction on medical records can help clinicians determine possible diagnoses and risk levels for chronic illness.[25]

Today's deep learning techniques also give us a much more accurate understanding of how diseases spread, giving us a better chance at prevention. Machine learning helps us model complex events, like the global COVID-19 pandemic, which require comprehensive epidemiological data sets, the development of novel interpretable models, and agent-based simulators to inform public health responses.[26]

Weather, environment, and climate change. Climate change is one of the greatest challenges currently facing humanity. Machine learning can help us better understand the weather and our environment, particularly to predict or forecast both everyday weather and climate disasters.

For weather and precipitation forecasting, computationally intensive physics-based models like the National Oceanic and Atmospheric Administration's

High-Resolution Rapid Refresh (HRRR) have long reigned supreme.[27] Machine learning–based forecasting systems can predict more accurately than the HRRR on short timescales, however, with better spatial resolution and faster forecast computations.[28]

For flood forecasting, neural networks can model river systems around the world (a technique called HydroNets), resulting in more accurate water-level predictions.[29] Utilizing this technology, authorities can send faster flood alerts, for example, to more than two hundred million people in India and Bangladesh.[30]

Machine learning also helps us better analyze satellite imagery. We can rapidly assess damage after a natural disaster (even with limited prior satellite imagery), understand the impact and extent of wildfires, and improve ecological and wildlife monitoring.[31]

Robotics. The physical world is messy, full of unexpected obstacles, slips, and breakages. This makes creating robots that can successfully operate in messy, real-world environments like kitchens, offices, and roadways quite challenging (industrial robotics has already had a significant impact on the world, operating in more-controlled environments like factory assembly lines). To hard-code or program real-world physical tasks, researchers need to anticipate all possible situations a robot might encounter. Machine learning efficiently trains robots to operate effectively in real-world environments through a combination of techniques like reinforcement learning, human demonstration, and natural language instruction. Machine learning also allows a more flexible, adaptable approach, in which robots can learn the best ways to engage in grasping or walking tasks rather than being locked into hard-coded assumptions.

Some interesting research techniques include automated reinforcement learning combined with long-range robotic navigation, teaching a robot to follow natural language instructions (in many languages!), and applying a zero-shot imitation learning framework to help robots better navigate simulated and real-world environments.[32]

Accessibility. It is easy to take for granted our ability to see a beautiful image, to hear a favorite song, or to speak with a loved one. Yet more than one billion people are not able to access the world in these ways. Machine learning improves accessibility by turning these signals – vision, hearing, speech – into other signals that can be well-managed by people with accessibility needs, enabling better access to the world around them. Some application examples include speech-to-text transcription, real-time transcriptions while someone is engaged in conversation, and applications that can help visually impaired users identify their surroundings.[33]

Individualized learning. Machine learning can also be used to create tools and applications that aid individualized learning. The benefits of this will be far reaching, and initial examples include early childhood reading coaching such as Google Read Along (formerly Bolo), which is helping children all over the world learn to

read in a variety of different languages,[34] and machine learning tools like Socratic that can help kids learn by giving them intuitive explanations and more detailed information about concepts they are grappling with, across a wide variety of subjects such as mathematics, chemistry, and literature.[35] Personalized learning backed by speech recognition, realistic speech output, and language understanding has the potential to improve educational outcomes across the world.

Computer-aided creativity. Deep learning algorithms show surprising abilities to transform images in sophisticated and creative ways, giving us the ability to easily create spaceships in the style of Monet or the Golden Gate Bridge in the style of Edvard Munch.[36] Via an algorithm for artistic style transfer (developed by machine learning researcher Leon Gatys and colleagues), a neural network can take a real-world image and an image of a painting and automatically render the real-world image in the style of the painter. DALL·E by OpenAI enables users to describe an image using text ("*armchairs in the shape of an avocado*" or "*a loft bedroom with a white bed next to a nightstand, with a fish tank standing beside the bed*") and generate images that have the properties expressed by the natural language description, making sophisticated tools for artists and other creators to quickly create images of what is in their head.[37]

Machine learning–powered tools are also helping musicians create in ways they never have before.[38] Moving beyond "technology," these new uses of computing can help anyone create new and unique sounds, rhythms, melodies, or even an entirely new musical instrument.

It is not hard to imagine future tools that can interactively help people create amazing representations of our mental imagery – "*Draw me a beach . . . no, I want it to be nighttime . . . with a full moon . . . and a mother giraffe with a baby next to a surfer coming out of the water*" – by just interactively talking to our computing assistants.

Important building blocks. Federated learning is a powerful machine learning approach that preserves user privacy while leveraging many distinct clients (such as mobile devices or organizations) to collaboratively train a model while keeping the training data decentralized.[39] This enables approaches that have superior privacy properties in large-scale learning systems.[40]

Researchers continue to push the state of the art in federated learning by developing adaptive learning algorithms, techniques for mimicking centralized algorithms in federated settings, substantial improvements in complimentary cryptographic protocols, and more.[41]

Transformers. Language has been at the heart of developing AI since the field began, given how ubiquitous language use and understanding is within our daily lives. Because language deals in symbols, it naturally prompted a symbolic approach to AI in the beginning. But over the years, AI researchers have come to realize that more statistical or pattern-based approaches yield better practical uses. The right types of deep learning can represent and manipulate the layered struc-

ture of language quite effectively for a variety of real-world tasks, from translating between languages to labeling images. Much of the work in this space from Google and elsewhere now relies on transformers, a particular style of neural network model originally developed for language problems (but with a growing body of evidence that they are also useful for images, videos, speech, protein folding, and a wide variety of other domains).[42]

There have been several interesting examples of transformers used in scientific settings, such as training on protein sequences to find representations encoding meaningful biological properties, protein generation via language modeling, bioBERT for text mining in biomedical data (with pretrained model and training code), embeddings of scientific text (with code), and medical question answering.[43] Computer scientists Maithra Raghu and Eric Schmidt have provided a comprehensive review of the ways in which deep learning has been used for scientific discovery.[44]

Machine learning for computer systems. Researchers are also applying machine learning to core computer science and computer systems problems themselves. This is an exciting virtuous cycle for machine learning and computing infrastructure research because it could accelerate the whole range of techniques that we apply to other fields. This trend is in fact spawning entire new conferences, such as MLSys.[45] Learning-based approaches are even being applied to database indices, learned sorting algorithms, compiler optimization, graph optimization, and memory allocation.[46]

Future of machine learning. A few interesting threads of research are occurring in the ML research community that will likely be even more interesting if combined.

First, work on sparsely activated models, such as the sparsely gated mixture of experts model, shows how to build very large capacity models in which just a portion of the model is "activated" for any given example (say, just two or three experts out of 2,048 experts).[47] The routing function in such models is trained simultaneously and jointly with the different experts, so that the routing function learns which experts are good at which sorts of examples, and the experts simultaneously learn to specialize for the characteristics of the stream of examples they are given. This is in contrast with most ML models today in which the whole model is activated for every example. Research scientist Ashish Vaswani and colleagues showed that such an approach is simultaneously about nine times more efficient for training, about 2.5 times more efficient for inference, and more accurate (+1 BLEU point, a relatively large improvement in accuracy for a language-translation task).[48]

Second, work on automated machine learning (AutoML), in which techniques such as neural architecture search or evolutionary architectural search can automatically learn effective structures and other aspects of machine learning mod-

els or components in order to optimize accuracy for a given task, often involves running many automated experiments, each of which may involve significant amounts of computation.[49]

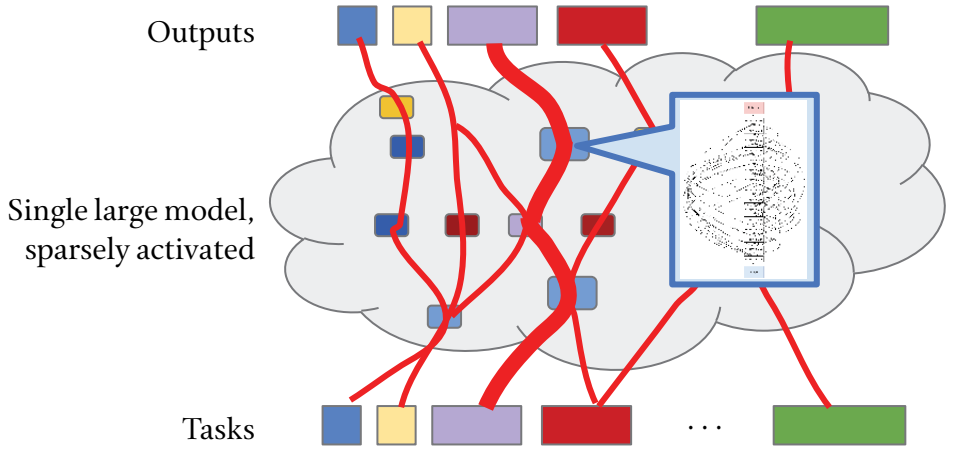
Third, multitask training at modest scales of a few to a few dozen related tasks, or transfer learning from a model trained on a large amount of data for a related task and then fine-tuned on a small amount of data for a new task, has been shown to be very effective in a wide variety of problems.[50] So far, most use of multitask machine learning is usually in the context of a single modality (such as all visual tasks or all textual tasks), although a few authors have considered multimodality settings as well.[51]

A particularly interesting research direction puts these three trends together, with a system running on large-scale ML accelerator hardware, with a goal of training a single model that can perform thousands or millions of tasks. Such a model might be made up of many different components of different structures, with the flow of data between examples being relatively dynamic on an example-by-example basis. The model might use techniques like the sparsely gated mixture of experts and learned routing in order to have a very large capacity model,[52] but one in which a given task or example only sparsely activates a small fraction of the total components in the system (and therefore keeps computational cost and power usage per training example or inference much lower). An interesting direction to explore would be to use dynamic and adaptive amounts of computation for different examples, so that "easy" examples use much less computation than "hard" examples (a relatively unusual property in the machine learning models of today). Figure 1 depicts such a system.

Each component might itself be running some AutoML-like architecture search in order to adapt the structure of the component to the kinds of data that are being routed to that component.[53] New tasks can leverage components trained on other tasks when that is useful. The hope is that through very large scale multitask learning, shared components, and learned routing, the model can very quickly learn to accomplish new tasks to a high level of accuracy, with relatively few examples for each new task (because the model is able to leverage the expertise and internal representations it has already developed in accomplishing other, related tasks).

Building a single machine learning system that can handle millions of tasks, and that can learn to successfully accomplish new tasks automatically, is a true grand challenge in the field of artificial intelligence and computer systems engineering. It will require expertise and advances in many areas, spanning machine learning algorithms, responsible AI topics such as fairness and interpretability, distributed systems, and computer architectures in order to push the field of artificial intelligence forward by building a system that can generalize to solve new tasks independently across the full range of application areas of machine learning.

Figure 1
A Multitask, Sparsely Activated Machine Learning Model



Note: This diagram depicts a design for a large, sparsely activated, multitask model. Each box in the model represents a component. Models for tasks develop by stitching together components, either using human-specified connection patterns or automatically learned connectivity. Each component might be running a small architectural search to adapt to the kinds of data that are being routed to it, and routing decisions making components decide which downstream components are best suited for a particular task or example, based on observed behavior. Source: Author's diagram, including Barret Zoph and Quoc V. Le, "Neural Architecture Search with Reinforcement Learning," arXiv (2016), Figure 7, 15, https://arxiv.org/abs/1611.01578.

Responsible AI development. While AI has the ability to help us in many facets of our lives, all researchers and practitioners should ensure that these approaches are developed responsibly – carefully reviewing issues of bias, fairness, privacy, and other social considerations on how these tools might behave and impact others – and work to address these considerations appropriately.

It is also important to document a clear set of principles to guide responsible development. In 2018, Google published a set of AI principles that guide the company's work in and use of AI.[54] The AI principles lay out important areas of consideration, including issues such as bias, safety, fairness, accountability, transparency, and privacy in machine learning systems. Other organizations and governments have followed this model by publishing their own principles around the use of AI in recent years. It is great to see more organizations publishing their own guidelines and I hope that this trend will continue until it is no longer a

trend but a standard by which all machine learning research and development is conducted.

Conclusions. The 2010s were truly a golden decade of deep learning research and progress. During this decade, the field made huge strides in some of the most difficult problem areas set out in the 1956 workshop that created the field of AI. Machines became capable of seeing, hearing, and understanding language in ways that early researchers had hoped for. The successes in these core areas enabled a huge range of progress in many scientific domains, enabled our smartphones to become much smarter, and generally opened our eyes to the possibilities of the future as we continue to make progress on creating more sophisticated and powerful deep learning models that help us with our daily lives. The future ahead of us is one in which we will all be more creative and capable thanks to the help provided by incredibly powerful machine learning systems. I cannot wait to see what the future holds!

AUTHOR'S NOTE

Alison Carroll, Heather Struntz, and Phyllis Bendell helped edit this manuscript and made many helpful suggestions for how to present much of the material.

ABOUT THE AUTHOR

Jeffrey Dean, a Fellow of the American Academy since 2016, is a Google Senior Fellow and Senior Vice President for Google Research at Google, Inc.; and Distinguished Fellow at the Stanford University Institute for Human-Centered Artificial Intelligence. He has published in such outlets as *Communications of the ACM*, *ACM Transactions on Computer Systems*, and *Transactions of the Association for Computational Linguistics*. His research papers can be found on Google Scholar at https://scholar.google.com/citations?user=NMS69lQAAAAJ.

ENDNOTES

[1] "Dartmouth Workshop," Wikipedia, last updated October 7, 2021, https://en.wikipedia.org/wiki/Dartmouth_workshop.

[2] "History of Artificial Intelligence," Wikipedia, last updated December 2, 2021, https://en.wikipedia.org/wiki/History_of_artificial_intelligence.

[3] "Cyc," Wikipedia, last updated October 21, 2021, https://en.wikipedia.org/wiki/Cyc.

[4] Jeffrey Dean, "Parallel Implementations of Neural Network Training: Two Back-Propagation Approaches" (senior thesis, University of Minnesota, 1990), https://drive.google.com/file/d/1I1fs4sczbCaACzA9XwxR3DiuXVtqmejL/view.

[5] Kyoung-Su Oh and Keechul Jung, "GPU Implementation of Neural Networks," *Pattern Recognition* 37 (6) (2004), https://www.sciencedirect.com/science/article/abs/pii/S0031320304000524.

[6] Rajat Raina, Anand Madhavan, and Andrew Y. Ng, "Large-Scale Deep Unsupervised Learning Using Graphics Processors," in *Proceedings of the 26th International Conference on Machine Learning* (Princeton, N.J.: International Machine Learning Society, 2009), http://robotics.stanford.edu/~ang/papers/icml09-LargeScaleUnsupervisedDeepLearningGPU.pdf.

[7] Jürgen Schmidhuber, "History of Computer Vision Contests Won by Deep CNNs on GPU," AI Blog, 2017, last updated 2021, https://people.idsia.ch/~juergen/computer-vision-contests-won-by-gpu-cnns.html.

[8] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton, "ImageNet Classification with Deep Convolution Neural Networks," in *Advances in Neural Information Processing Systems 25 (NIPS 2012)*, ed. F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger (Neural Information Processing Systems Foundation, 2012), https://papers.nips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf; Geoffrey Hinton, Li Deng, Dong Yu, et al., "Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups," *IEEE Signal Processing Magazine* 29 (6) (2012), https://ieeexplore.ieee.org/document/6296526; Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean, "Efficient Estimation of Word Representations in Vector Space," arXiv (2013), https://arxiv.org/abs/1301.3781; and Ilya Sutskever, Oriol Vinyals, and Quoc V. Le, "Sequence to Sequence Learning with Neural Networks," arXiv (2014), https://arxiv.org/abs/1409.3215.

[9] Jeffrey Dean, Greg S. Corrado, Rajat Monga, et al., "Large Scale Distributed Deep Networks" (Mountain View, Calif.: Google, Inc., 2012), https://static.googleusercontent.com/media/research.google.com/en//archive/large_deep_networks_nips2012.pdf.

[10] Quoc V. Le, Marc'Aurelio Ranzato, Rajat Monga, et al., "Building High-Level Features Using Large Scale Unsupervised Learning," in *Proceedings of the 29th International Conference on Machine Learning* (Princeton, N.J.: International Machine Learning Society, 2012), https://static.googleusercontent.com/media/research.google.com/en//archive/unsupervised_icml2012.pdf.

[11] Norman P. Jouppi, Cliff Young, Nishant Patil, et al., "In-Datacenter Performance Analysis of a Tensor Processing Unit," in *ISCA '17: Proceedings of the 44th Annual International Symposium on Computer Architecture* (New York: Association for Computing Machinery, 2017), https://dl.acm.org/doi/10.1145/3079856.3080246.

[12] Norman P. Jouppi, Doe Hyun Yoon, George Kurian, et al., "A Domain-Specific Supercomputer for Training Deep Neural Networks," *Communications of the ACM* 63 (7) (2020), https://cacm.acm.org/magazines/2020/7/245702-a-domain-specific-supercomputer-for-training-deep-neural-networks/fulltext.

[13] Ronan Collobert, Samy Bengio, and Johnny Mariéthoz, "Torch: A Modular Machine Learning Software Library," IDIAP Research Report 02-46 (Martigny, Switzerland: Dalle Molle Institute for Perceptual Artificial Intelligence, 2002), https://infoscience.epfl.ch/record/82802/files/rr02-46.pdf.

[14] James Bergstra, Oliver Breuleux, Frédéric Bastien, et al., "Theano: A CPU and GPU Math Compiler in Python," in *Proceedings of the 9th Python in Science Conference (SciPy 2010)* (Austin: Python in Science Conference, 2010), http://conference.scipy.org/proceedings/scipy2010/pdfs/bergstra.pdf.

[15] Dean et al., "Large Scale Distributed Deep Networks"; and "Caffe," Berkeley Artificial Intelligence Research, https://caffe.berkeleyvision.org/.

[16] TensorFlow, https://www.tensorflow.org/.

[17] "TensorFlow: A System for Large-Scale Machine Learning," Usenix, https://www.usenix.org/conference/osdi16/technical-sessions/presentation/abadi.

[18] Adam Paszke, Sam Gross, Francisco Massa, et al., "PyTorch: An Imperative Style, High-Performance Deep Learning Library," arXiv (2019), https://arxiv.org/abs/1912.01703.

[19] Matthew Johnson, Peter Hawkins, Jake Vanderplas, et al., "Jax," GitHub, last updated December 15, 2021, http://github.com/google/jax; and "XLA: Optimizing Compiler for Machine Learning," TensorFlow, last updated December 2, 2021, https://www.tensorflow.org/xla.

[20] "Machine Learning," arXiv, https://arxiv.org/list/cs.LG/recent.

[21] Michal Januszewski, "Releasing the Drosophila Hemibrain Connectome–The Largest Synapse-Resolution Map of Brain Connectivity," Google AI Blog, January 22, 2020, https://ai.googleblog.com/2020/01/releasing-drosophila-hemibrain.html.

[22] Janelia FlyEM Hemibrain Dataset Information, Hemibrain Neuroglancer Demo, https://tinyurl.com/25cjs3uk.

[23] Saud H. AlDubayan, Jake R. Conway, Sabrina Y. Camp, et al., "Detection of Pathogenic Variants with Germline Genetic Testing Using Deep Learning vs. Standard Methods in Patients with Prostate Cancer and Melanoma," *JAMA* 324 (19) (2020), https://jamanetwork.com/journals/jama/article-abstract/2772962?guestAccessKey=39889aad-2894-4380-b869-5704ed2f9f6b.

[24] Shravya Shetty and Daniel Tse, "Using AI to Improve Breast Cancer Screening," The Keyword, January 1, 2020, https://blog.google/technology/health/improving-breast-cancer-screening/; and Yuan Liu, Ayush Jain, Clara Eng, et al., A Deep Learning System for Differential Diagnosis of Skin Diseases," *Nature Medicine* 26 (2020), https://www.nature.com/articles/s41591-020-0842-3.

[25] Alvin Rajkomar and Eyal Oren, "Deep Learning for Electronic Health Records," Google AI Blog, May 8, 2018, https://ai.googleblog.com/2018/05/deep-learning-for-electronic-health.html.

[26] "Covid-19-open-data," GitHub, updated December 1, 2021, https://github.com/GoogleCloudPlatform/covid-19-open-data; Sercan Arik, Chun-Liang Li, Jinsung Yoon, et al., "Interpretable Sequence Learning for Covid-19 Forecasting," in *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, ed. H. Larochelle, M. Ranzato, R. Hadsell, et al. (Neural Information Processing Systems Foundation, 2020), https://research.google/pubs/pub49500/; and "Agent-based-epidemic-sim," GitHub, updated September 28, 2021, https://github.com/google-research/agent-based-epidemic-sim.

[27] "The High-Resolution Rapid Refresh (HRRR)," Global Systems Laboratory, U.S. Department of Commerce, https://rapidrefresh.noaa.gov/hrrr/.

[28] Jason Hickey, "Using Machine Learning to 'Nowcast' Precipitation in High Resolution," Google AI Blog, January 13, 2020, https://ai.googleblog.com/2020/01/using-machine-learning-to-nowcast.html.

[29] Sella Nevo, "The Technology Behind Our Recent Improvements in Flood Forecasting," Google AI Blog, September 3, 2020, https://ai.googleblog.com/2020/09/the-technology-behind-our-recent.html.

[30] Yossi Matias, "A Big Step for Flood Forecasts in India and Bangladesh," The Keyword, September 1, 2020, https://blog.google/technology/ai/flood-forecasts-india-bangladesh/.

[31] Joseph Xu and Pranav Khaitan, "Machine Learning-Based Damage Assessment for Disaster Relief," Google AI Blog, June 16, 2020, https://ai.googleblog.com/2020/06/machine-learning-based-damage.html; Jihyeon Lee, Joseph Z. Xu, Kihyuk Sohn, et al., "Assessing Post-Disaster Damage from Satellite Imagery Using Semi-Supervised Learning Techniques," arXiv (2011), https://arxiv.org/abs/2011.14004; Yossi Matias, "Mapping Wildfires with the Power of Satellite Data," The Keyword, August 20, 2020, https://blog.google/products/search/mapping-wildfires-with-satellite-data/; and Sara Beery and Jonathan Huang, "Leveraging Temporal Context for Object Detection," Google AI Blog, June 26, 2020, https://ai.googleblog.com/2020/06/leveraging-temporal-context-for-object.html.

[32] Aleksandra Faust and Anthony Francis, "Long-Range Robotic Navigation via Automated Reinforcement Learning," Google AI Blog, February 28, 2019, https://ai.googleblog.com/2019/02/long-range-robotic-navigation-via.html; Corey Lynch and Pierre Sermanet, "Language Conditioned Imitation Learning over Unstructured Data" (2020), https://language-play.github.io/; and Xinlei Pan, Tingnan Zhang, Brian Ichter, et al., "Zero-Shot Imitation Learning from Demonstrations for Legged Robot Visual Navigation," *ICRA* (2020), https://research.google/pubs/pub48968/.

[33] Julie Cattiau, "How AI Can Improve Products for People with Impaired Speech," The Keyword, May 7, 2019, https://www.blog.google/outreach-initiatives/accessibility/impaired-speech-recognition/; Sagar Savla, "Real-Time Continuous Transcription with Live Transcribe," Google AI Blog, February 4, 2019, https://ai.googleblog.com/2019/02/real-time-continuous-transcription-with.html; and "Lookout by Google," Google Play, https://play.google.com/store/apps/details?id=com.google.android.apps.accessibility.reveal.

[34] Avni Shah, "Making Learning to Read Accessible and Fun with Bolo," The Keyword, September 9, 2019, https://www.blog.google/technology/ai/bolo-literacy/.

[35] Shreyans Bhansali, "When Students Get Stuck, Socratic Can Help," The Keyword, August 15, 2019, https://www.blog.google/outreach-initiatives/education/socratic-by-google/.

[36] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge, "A Neural Algorithm of Artistic Style," arXiv (2015), https://arxiv.org/abs/1508.06576; and Vincent Dumoulin, Jonathon Shlens, and Manjunath Kudlur, "Supercharging Style Transfer," Google AI Blog, October 26, 2016, https://ai.googleblog.com/2016/10/supercharging-style-transfer.html.

[37] "DALL·E: Creating Images from Text," OpenAI, https://openai.com/blog/dall-e/.

[38] "Magenta," Google Research Team, https://research.google/teams/brain/magenta/.

[39] Jakub Konečný, Brendan McMahan, and Daniel Ramage, "Federated Optimization: Distributed Optimization beyond the Datacenter," arXiv (2015), https://arxiv.org/abs/1511.03575.

[40] Keith Bonawitz, Hubert Eichner, Wolfgang Grieskamp, et al., "Towards Federated Learning at Scale: System Design," arXiv (2019), https://arxiv.org/abs/1902.01046. For a less technical description on how the technology works, you can also check out this online comic description: "Federated Learning: Building Better Products with On-Device Data and Privacy by Default," Google AI, https://federated.withgoogle.com/.

[41] Sashank Reddi, Zachary Charles, Manzil Zaheer, et al., "Adaptive Federated Optimization," arXiv (2020), https://arxiv.org/abs/2003.00295; Sai Praneeth Karimireddy, Martin Jaggi, Satyen Kale, et al., "Mime: Mimicking Centralized Stochastic Algorithms in Federated Learning," arXiv (2020), https://arxiv.org/abs/2008.03606; and James Bell, K. A. Bonawitz, Adrià Gascón, et al., "Cryptology ePrint Archive: Report 2020/704," in *CCS '20: Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security* (New York: Association for Computing Machinery, 2020), https://eprint.iacr.org/2020/704.

[42] Ashish Vaswani, Noam Shazeer, Niki Parmar, et al., "Attention Is All You Need," arXiv (2017), https://arxiv.org/abs/1706.03762; Jakob Uszkoreit, "Transformer: A Novel Neural Network Architecture for Language Understanding," Google AI Blog, August 31, 2017, https://ai.googleblog.com/2017/08/transformer-novel-neural-network.html; Neil Houlsby and Dirk Weissenborn, "Transformers for Image Recognition at Scale," Google AI Blog, December 3, 2020, https://ai.googleblog.com/2020/12/transformers-for-image-recognition-at.html; Dirk Weissenborn, Oscar Täckström, and Jakob Uszkoreit, "Scaling Autoregressive Video Models," arXiv (2019), https://arxiv.org/abs/1906.02634; Anmol Gulati, James Qin, Chung-Cheng Chiu, et al., "Conformer: Convolution-Augmented Transformer for Speech Recognition," arXiv (2020), https://arxiv.org/abs/2005.08100; and AlphaFold Team, "AlphaFold: A Solution to a 50-Year-Old Grand Challenge in Biology," DeepMind, November 30, 2020, https://deepmind.com/blog/article/alphafold-a-solution-to-a-50-year-old-grand-challenge-in-biology.

[43] Kyle Lo, Iz Beltagy, Arman Cohan, et al., "Scibert," GitHub, last updated June 14, 2020, https://github.com/allenai/scibert.

[44] Maithra Raghu and Eric Schmidt, "A Survey of Deep Learning for Scientific Discovery," arXiv (2020), https://arxiv.org/abs/2003.11755.

[45] "MLSys–2020," Fifth Conference on Machine Learning and Systems, Santa Clara, California, April 11–14, 2020, https://mlsys.org/.

[46] Tim Kraska, Alex Beutel, Ed H. Chi, et al., "The Case for Learned Index Structures," arXiv (2017), https://arxiv.org/abs/1712.01208; Ani Kristo, Kapil Vaidya, Ugur Çetintemel, et al., "The Case for a Learned Sorting Algorithm," in *SIGMOD '20: Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data* (New York: Association for Computing Machinery, 2020), https://dl.acm.org/doi/abs/10.1145/3318464.3389752; Samuel J. Kaufman, Phitchaya Mangpo Phothilimthana, Yanqi Zhou, et al., "A Learned Performance Model for Tensor Processing Units," arXiv (2020), https://arxiv.org/abs/2008.01040; Yanqi Zhou and Sudip Roy, "End-to-End, Transferable Deep RL for Graph Optimization," Google AI Blog, December 17, 2020, https://ai.googleblog.com/2020/12/end-to-end-transferable-deep-rl-for.html; and Martin Maas, David G. Andersen, Michael Isard, et al., "Learning-Based Memory Allocation for C++ Server

Workloads," *Proceedings of the 25th ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS)* (New York: Association for Computing Machinery, 2020), https://research.google/pubs/pub49008/.

47 Vaswani et al., "Attention Is All You Need."

48 Ibid., Table 4.

49 Barret Zoph and Quoc V. Le, "Neural Architecture Search with Reinforcement Learning," arXiv (2016), https://arxiv.org/abs/1611.01578; Hieu Pham, Melody Guan, Barret Zoph, et al., "Efficient Neural Architecture Search via Parameters Sharing," *Proceedings of Machine Learning Research* 80 (2018), http://proceedings.mlr.press/v80/pham18a.html; Adam Gaier and David Ha, "Weight Agnostic Neural Networks," arXiv (2019), https://arxiv.org/abs/1906.04358; and Esteban Real, Sherry Moore, Andrew Selle, et al., "Large-Scale Evolution of Image Classifiers," arXiv (2017), https://arxiv.org/abs/1703.01041.

50 Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," arXiv (2018), https://arxiv.org/abs/1810.04805.

51 Carl Doersch and Andrew Zisserman, "Multi-Task Self-Supervised Visual Learning," arXiv (2017), https://arxiv.org/abs/1708.07860; and Sebastian Ruder, "An Overview of Multi-Task Learning in Deep Neural Networks," arXiv (2017), https://arxiv.org/abs/1706.05098.

52 Vaswani et al., "Attention Is All You Need."

53 Pham et al., "Efficient Neural Architecture Search via Parameters Sharing."

54 "Artificial Intelligence at Google: Our Principles," Google AI, https://ai.google/principles/.

I Do Not Think It Means What You Think It Means: Artificial Intelligence, Cognitive Work & Scale

Kevin Scott

Over the past decade, AI technologies have advanced by leaps and bounds. Progress has been so fast, voluminous, and varied that it can be a challenge even for experts to make sense of it all. In this essay, I propose a framework for thinking about AI systems, specifically the idea that they are ultimately tools developed by humans to help other humans perform an increasing breadth of their cognitive work. Our AI systems for assisting us with our cognitive work have become more capable and general over the past few years. This is in part due to a confluence of novel AI algorithms and the availability of massive amounts of data and compute. From this, researchers and engineers have been able to construct large, general models that serve as flexible and powerful building blocks that can be composed with other software to drive breakthroughs in the natural and physical sciences, to solve hard optimization and strategy problems, to perform perception tasks, and even to assist with complex cognitive tasks like coding.

When I say the word "work," what do I mean? In the mornings, when I tell my children that "I'm going to work," they understand that I am about to get into a car, drive to my office, and, for the rest of the day, do a set of things alongside my coworkers for an employer who pays my salary. When I tell my wife that "I'm going to work in the shop for a while," she understands that I am headed to my workshop where I will use a variety of tools that I hold dear to tinker around on personal projects. When I say that "I'm going to work in the garden" or "I'm going to work on this essay," the people to whom I am speaking almost always understand what I mean. Work in all these contexts means me, a human being, applying effort to achieve some effect. In these contexts, we all have some shared understanding of what the applied efforts entail, and why the effects are worth achieving.

In the late eighteenth century, accelerating into the nineteenth and twentieth centuries, individual members of society had cause to think about work in new ways. As society industrialized and humans devised new ways to use machines to

© 2022 by Kevin Scott
Published under a Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) license
https://doi.org/10.1162/DAED_a_01901

do work, nearly every aspect of human life changed. As these machines became increasingly complex, and as we began to use them to perform types of work that had previously been performed through a combination of human labor and less powerful tools, we needed new language and new scientific, technical, and social shared understandings for these new forms of machine-assisted work.

Driven by the intellectual and industrial revolutions of the period, by the last half of the nineteenth century, scientists like Nicolas Léonard Sadi Carnot, James Prescott Joule, Rudolf Clausius, Lord Kelvin, James Maxwell, Ludwig Boltzmann, and others had given us a simple but powerful definition of work – weight lifted through a height – and a rich scientific theory – thermodynamics – that helped us better understand not just the natural world, but how to better engineer, build, and direct the new forms of machine work shaping society. That nineteenth-century scientific definition of work is very much relevant today, but it is characteristic of its time. The work that it defines is physical. Understanding the nature of physical work was and is necessary to understand the machinery of the universe and was essential in constructing an industrial society.

When I get in my car, drive to my office, and do things alongside my coworkers, "weight lifted through a height" is perhaps not the most relevant definition of the work that I, and many others, do every day. I meet with people. I listen. I coach and mentor. I attempt to make a very small number of meaningful decisions. I read and digest information. I think. I imagine. I code. I write. With all these efforts, the effects that I am trying to achieve are the solutions of problems. For me those might be: Can we use our AI supercomputers to make molecular dynamics simulations go much faster so that we can solve a more interesting set of problems in biology? How can we make sure that our next machine learning model does not produce adverse effects? Can we work around firmware issues to prevent a compute shortage in our AI training clusters? Can I understand enough of what a coworker is trying to achieve to meaningfully assist them? The interesting thing about all these problems and their solutions is that the work required to solve them is almost entirely cognitive.

If you are reading this essay, I would wager that you earn some or all of your living doing cognitive work. Perhaps, if you made a full accounting of your work time, you would discover that, if not most of the effort that you exert in your work, then a majority of the effects that you produce are more of the mind than the body. I am not arguing that our bodies are mere instruments of the mind. And I am certainly not arguing that one form of work is superior to another. I am attempting to make a more prosaic assertion: I am a knowledge worker; and you may be, too. Moreover, even though we understand the nature of our work well enough to do it, and more of us are earning our living this way with each passing year, we have not yet crisply defined what cognitive work is nor how to measure

it. As AI technologies become more capable, and as we use them to do more things that are inarguably cognitive work, this lack of a foundational definition makes it increasingly difficult to predict and engineer the changes that machines will bring to cognitive work in the coming years. Will AI become yet another instrument or tool that we use to express our humanity and creativity, that allows us to better explore and understand ourselves and the world around us, and that evolves the nature of work once again just as the machines of the industrial revolution have done over the past two centuries? Or will AI become something else?

Wikipedia dodges defining what knowledge work is by defining the knowledge worker instead as someone whose main capital is knowledge. The knowledge worker entry then lists examples: "programmers, physicians, pharmacists, architects, engineers, scientists, design thinkers, public accountants, lawyers, editors, and academics, whose job is to '*think for a living*.'"[1] Not bad. But not good enough to build a theory of cognitive work as useful as thermodynamics was for physical work.

Since the middle of the twentieth century, we have had mathematician Claude Shannon's quantification of information and an information theory with connections to and, in some respects, directly inspired by classical thermodynamics.[2] Intuitively, it seems safe to say that information is the precursor to knowledge. In some sense, building the bridge from the rigor of information theory to a useful theory of cognitive work has been one of the great challenges facing the discipline of AI since its founding in the summer of 1956. You can well imagine that our ancestors faced a similar quandary in the eighteenth and nineteenth centuries as they architected the industrialization of society. Sometimes the machines came before we really understood why they worked and the best way to build them, much less the complex network of social implications their construction and use entailed. But our ancestors built those machines anyway because it was blindingly obvious why they were useful.

In 2022, we have more clues about what a theory of cognitive work might be, although the theory itself may not be a new one. Of the ten attendees of the 1956 Dartmouth Summer Research Conference on Artificial Intelligence, which coined the term *artificial intelligence* and helped to establish AI as a discipline, Ray Solomonoff's name is less well-known than Marvin Minsky, John McCarthy, or Claude Shannon. Even though the subdiscipline of AI called machine learning has only in the past two decades taken over as the primary thrust of AI research and commercial activity, from the beginning, Solomonoff envisioned machine systems that could use probability and data to learn to solve complex problems.

Perhaps the most important of Solomonoff's insights was his theory of inductive inference. This theory is in some ways a resolution of tension between two ancient ideas: Occam's razor and Epicurus's principle of multiple explanations. We are probably all familiar with Occam's razor, which states that when faced with

a choice between multiple consistent explanations of an observed phenomenon, we should choose the simplest. Epicurus's principle, on the other hand, states that we should consider all consistent explanations.

Solomonoff's resolution, while mathematically quite sophisticated, is a relatively simple idea. You formulate the explanations of observable phenomena as programs for an abstract computing device, specifically a universal Turing machine.[3] The shorter a program is, the more concise it is at explaining observed phenomena. We can now use this conciseness as a precise measure of simplicity for Occam's razor. We then use the tools of Bayesian probability and a universal prior to compute the posterior probability of the range of computable explanations for any observed phenomenon.

When we train modern machine learning models, to be clear, we are not performing Solomonoff induction, which Solomonoff himself proved to be uncomputable. Regardless, Solomonoff induction is an interesting framework for thinking about cognitive work given that it is complete, at least over the universe of computable explanations.[4] Although I am biased by my computer science training, I would argue that it is not hard to imagine how you could explain almost any observable phenomenon by at least some arbitrarily long program. The beauty of Solomonoff induction is that, to quote Ilya Sutskever, chief scientist of OpenAI, "compression equals generalization." An incomprehensibly long explanation of a single phenomenon is nowhere near as powerful as a single concise description of many phenomena. Solomonoff induction gives us a framework for thinking very precisely about exactly this.

I may have just invoked too little theoretical computer science to frustrate the real theoretical computer scientists, and too much to frustrate everyone else, with the question still lingering: how does this help us understand cognitive work? Let us step back a moment to the work that we all do as knowledge workers. Much of our work involves the use of a bunch of cognitive tools that humans have developed over millennia, and frameworks for refining and composing these tools with one another that help us solve problems orders and orders of magnitude more complex than our ancestors could, even though biologically we are most certainly not orders of magnitude smarter. Our ability to refine these cognitive tools, to rigorously ensure that they work, and then to compose them may very well be the human version of "compression is generalization," the way for us to do more even though we likely have no more real cognitive capacity than the ancients.

Take two of these tools I am guessing that many of us use to do our work: mathematics and the scientific method. The modern body of mathematics that we learn in high school and university, and increasingly the computational tools that we use to support our mathematical activities, lets us reason about phenom-

ena we can neither see, touch, nor otherwise sense. Perhaps more important, it allows us to make predictions and reason about phenomena that have never actually occurred. With millennia-old mathematics, our ancestors could design aqueducts that supported sophisticated ancient civilizations by allowing them to move water around for irrigation, drinking, and sanitation. With twentieth- and twenty-first-century mathematics and computation, we can design lithographic structures on silicon wafers that move electrons around with near atomic-level precision. We carry devices made with these silicon artifacts in our pockets and backpacks that give us a way to connect and communicate with billions of other humans, access the world's knowledge, create our work, and engage in almost any form of commerce imaginable. To get from aqueducts to microprocessors, we have had to build a whole modern cognitive architecture composed of layers upon layers of cognitive tools that we and our predecessors have contributed to.

When I stop to think hard about the tools that I use to do my work, they do feel like an amazing compression algorithm that lets me get more mileage out of the brain I was born with. In computer science, this effect is hard to miss. The programs that I wrote as a young computing professional were longer and accomplished far less than the ones I write today. And the margin by which a line of code has become more powerful is far greater than the productivity I have gained through polishing my programming skills over the years. The tools that are available to me now are orders of magnitude more powerful than they were when I began coding in the 1980s. Moreover, whether you are an engineer, a scientist, a writer, or an artist, what has become clear over the past handful of years is that the AI systems that we are building today will likely have an equally momentous impact on the cognitive work that we are all able to do in the future.

In the same way that an engineer might assemble metal alloys, hydraulic pistons, electric motors, shafts, bearings, and electronics into a machine that performs mechanical work, like a forklift, engineers of AI systems increasingly rely upon deep neural networks (DNNs) to build software systems capable of performing cognitive work. In a real sense, the widespread use of DNNs today is made possible by large amounts of data and compute needed to train them. In 2009, machine learning scholar Andrew Ng and his colleagues at Stanford proposed the use of graphics processing units (GPUs) – devices capable of quickly and efficiently performing the sorts of arithmetic necessary for creating realistic video games – for training DNNs.[5] While Ng did not invent the DNN, his innovative use of the computational power of GPUs to train them helped to bring about a new age of machine learning with the DNN as its most powerful building block.

Over the past decade or so, the amount of compute used to train the DNN building blocks of our AI tools for cognitive work has increased exponentially. In 2018, OpenAI scientists noted that from 2012 to 2018, the amount of compute used

in the largest AI training runs had increased by a factor of three hundred thousand.[6] Why? In a world of diminishing returns from Moore's law, it certainly is not because compute is cheap. These investments only make sense insofar as scale makes DNNs better building blocks for doing cognitive work. And arguably they have, in two notable ways.

In the first half of the nineteenth century, mechanical engineer Claude-Louis Navier and physicist George Gabriel Stokes developed a set of partial differential equations to describe the motion of viscous fluids. The Navier-Stokes flow equations are, in my opinion, among the most beautiful in all of mathematics. They very concisely describe an enormous range of phenomena in hydraulics, aeronautics, oceanography, and atmospheric sciences. They inform everything from the design of the pipes carrying water to our homes, to the design of the aircraft that take us on holiday, to the weather forecasts we use to plan our days. The problem with these equations is that, when used to model extremely complex physical objects or environments, they can become extraordinarily expensive to solve. Prior to the advent of computers and numerical solvers for partial differential equations (PDEs), one could only model relatively simple systems with high fidelity, or complex systems only with simplifying assumptions to make the calculations feasible. Even now with extremely powerful supercomputers, certain problems that could benefit from high-fidelity solutions to Navier-Stokes are computationally infeasible.

That is, until recent work by a team of computer scientists at Caltech. Zongyi Li and colleagues devised a way to use deep neural networks to solve the Navier-Stokes PDEs up to three orders of magnitude faster, under some circumstances, than prior state-of-the-art solvers.[7] In my graduate research, I was often happy to improve the performance of a system by 5 percent. One thousand times more performance is, to torture an overused word, incredible.

The pattern that Li and his colleagues employed is one that is becoming increasingly widespread in the sciences. This is the first notable way in which models trained with large amounts of compute are becoming better building blocks for cognitive work. With an abundance of compute, DNNs can be trained using accurate but slow simulators or solvers for numerical, combinatorial, or even symbolic problems to encode something about the structure of a problem domain that we have yet to be able to model in other ways, such as through mathematics, or heuristics, or code. These DNNs can then be used to solve problems, allowing scientists to approach their work in new ways. Sometimes these techniques may make expensive things quicker or cheaper so that more people can solve more problems. Sometimes they may mean creating the ability to tackle problems so large or complex that they were previously impossible to solve. And the better news is that it seems as if this pattern is widely applicable and just beginning to be widely adopted. There is much to look forward to in the years to come.

The second way that scale is allowing us to construct better building blocks for performing cognitive work involves the use of self-supervised learning for building deep neural networks that behave as building blocks or platforms for a wide range of uses.

Before we dive into an explanation of self-supervised models, it is useful to understand a bit about supervised models, which drove much of the progress in the early years of the DNN boom. The first decade or more of machine learning systems that I built were all supervised. If you wanted to train a model to predict when someone was going to click on an ad, whether a piece of email was spam, or whether a picture contains an image of a kitten, you had to provide the supervised learning algorithms with lots and lots of examples of both good and bad ad clicks, spammy and nonspammy emails, or pictures with or without kittens in them. Providing those examples and counterexamples is an exercise called labeling and is time consuming and expensive given the volume of labeled training data required to achieve good performance.

For those of us following the field of machine learning closely, the last several years have brought extraordinary progress in solving problems related to human perception (recognizing the objects in images or the words spoken to a device), strategic game playing (beating the best human players at Go or Dota), and, most recently, in natural language understanding. The progress in natural language understanding began to accelerate in 2018 with the publication of a paper by Jacob Devlin, a software engineer at Google, which introduced the notion of *pretraining* for language models.[8] By now, this will feel like a familiar pattern. BERT, RoBERTa, DeBERTa, and other models use a set of techniques to learn the structure of natural language in a process that researchers in the field call pretraining.[9] Pretraining in these language models, like many of the most powerful contemporary deep learning systems, is self-supervised. In other words, the models learn without direct human supervision.

Once pretrained, a model, with the things it has learned about language structure, can be used to solve a wide range of problems in natural language processing. In many cases, a pretrained model needs to be fine-tuned to a particular task with some supervision. In some cases, the pretrained model itself is good enough. For instance, researchers at the Allen Institute for Artificial Intelligence used BERT in a test-taking system they had built called Aristo that was able to score higher than 90 percent on the multiple-choice component of the New York Regents eighth-grade science exam, and exceeded 83 percent on the twelfth-grade test.[10] My colleagues at Microsoft Research used their DeBERTa model to, for the first time, surpass the human baseline on the SuperGLUE benchmark, which entails solving nontrivial natural language problems, such as processing a complicated passage of text and then answering true or false questions about the passage, or

resolving the referent of a pronoun in an ambiguous passage of text. The best natural language models are now able to exceed expert human performance on these benchmarks.

When examining these systems, it is always important to ask: are these models capable of doing what they do because they have superhumanly big memories from which they recall the answers to problems someone else has solved, or have they compressed what they have seen in a way that lets them generalize solutions of problems no one has ever solved before? While there is ongoing debate about what, if anything, our contemporary self-supervised language models are "learning,"[11] for both those systems as well as those in which the data-fueling model learning is generated in simulation, it does seem that large data and compute are allowing us to encode useful things about problem domains that no human has previously encountered.

Perhaps the two most impressive recent illustrations of how large models trained with large compute can produce interesting results are OpenAI's Generative Pre-trained Transformer 3 (GPT-3) and Codex models.[12] At the time of its release, GPT-3 was ten times larger than the largest nonsparse language model. There are many benefits to scale, although perhaps the two most important are: when properly trained, larger models tend to have better performance on the same task than smaller models; and larger models tend to be useful in a broader range of tasks, either with fine-tuning or not, than smaller models. Because GPT-3 is useful on a broad range of tasks with little or no additional fine-tuning, it has been possible to offer an application programming interface to developers to allow them to probe the utility of the model for the problems they are interested in solving.

One of the biggest surprises of the GPT-3 model is that it generalized something about the structure of computer programming languages that allowed it to synthesize code that did not exist in its training data. This realization led to OpenAI fine-tuning a model for computer code called Codex, and in collaboration with GitHub, developing a coding assistant product called Copilot that can write code from natural language prompts. As the Codex model and the Copilot product get better, they will not only assist programmers with their cognitive work, but may also lower the barrier to entry to programming itself. Given that Codex and Copilot work by allowing humans to describe in natural language an effect they would like accomplished through code, the task of coding may become more approachable to many, many more people.

This ability to train on one set of data and to transfer what is learned to a broad range of tasks is called transfer learning. Transfer learning, perhaps more than anything else over the next few years, is likely to accelerate our progress on AI. It allows us to think about models as reusable building blocks, what I call *platform models*, and researchers at Stanford are calling *foundation*

models.[13] Moreover, based on the trends of the past few years, for transfer learning to work better, we will need bigger and more sophisticated models, which in turn will require more training compute.

AI systems designed to assist us with our cognitive work will no doubt continue to surprise us. I have been surprised so many times over the past two decades by what AI scientists and researchers have been able to accomplish that I have learned to heed the second half of Arthur C. Clarke's first law: When a distinguished but elderly scientist states that something is possible, they are almost certainly right. When they state that something is impossible, they are very probably wrong. Somewhere in the surprises of the future that await us, I am looking forward to systems that can help me to write my code, to sharpen my writing, to help me better manage the deluge of information I crave, and to assist me with the art and artifacts I make in my workshop. Hopefully, as our eighteenth- and nineteenth-century forebears did with physical work, we will also sharpen our definitions of cognitive work, develop new mechanisms for measuring it, and get better at constructing AI building blocks and tools to help us with these tasks. But more than anything, I look forward to what happens when folks who are more imaginative and creative than I am are able to incorporate new AI-based cognitive tools into their work, to make things that awe and inspire, and to solve those vexing problems that face society as we race forward to an ever more complicated future.

ABOUT THE AUTHOR

Kevin Scott is Chief Technology Officer and Executive Vice President of Technology and Research at Microsoft. He is the author of *Reprogramming the American Dream: From Rural America to Silicon Valley – Making AI Serve Us All* (2020), is a coinventor on multiple patents, and is host of the podcast *Behind the Tech*.

ENDNOTES

1 "Knowledge Worker," Wikipedia, last updated October 5, 2021, https://en.wikipedia.org/wiki/Knowledge_worker (accessed January 27, 2022; emphasis added).

2 Claude Shannon and Warren Weaver, *A Mathematical Theory of Communication* (Champaign: University of Illinois Press, 1949).

3 Alan Turing, "On Computable Numbers, with an Application to the Entscheidungsproblem," *Proceedings of the London Mathematical Society* (1936): 230–265.

4 I am deliberately avoiding the use of the term *intelligence* given that its use is so burdened with imprecision and poor analogy.

[5] Rajat Raina, Anand Madhavan, and Andrew Y. Ng, "Large-Scale Deep Unsupervised Learning Using Graphics Processors," in *Proceedings of the 26th International Conference on Machine Learning* (New York: Association for Computing Machinery, 2009).

[6] Dario Amodei and Danny Hernandez, "AI and Compute," OpenAI, May 16, 2018, https://openai.com/blog/ai-and-compute/.

[7] Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, et al., "Fourier Neural Operator for Parametric Partial Differential Equations," presented at the 9th International Conference on Learning Representations, virtual event, May 3–7, 2021.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," arXiv (2020), https://arxiv.org/pdf/1810.04805.pdf.

[9] Yinhan Liu, Myle Ott, Naman Goyal, et al., "RoBERTa: A Robustly Optimized BERT Pretraining Approach," arXiv (2019), https://arxiv.org/pdf/1907.11692.pdf; and Pencheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen, "DeBERTa: Decoding-Enhanced BERT with Disentangled Attention," arXiv (2020), https://arxiv.org/pdf/2006.03654.pdf.

[10] Peter Clark, Oren Etzioni, Daniel Khashabi, Tushar Khot, et al., "From 'F' to 'A' on the N.Y. Regents Science Exams: An Overview of the Aristo Project," arXiv (2021), https://arxiv.org/pdf/1909.01958.pdf.

[11] Emily Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell, "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜" in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (New York: Association of Computing Machinery, 2021).

[12] Tom Brown, Benjamin Mann, Nick Ryder, et al., "Language Models Are Few-Shot Learners," arXiv (2020), https://arxiv.org/pdf/2005.14165v4.pdf; and Mark Chen, Jerry Tworek, and Heewoo Jun, "Evaluating Large Language Models Trained on Code," arXiv (2021), https://arxiv.org/abs/2107.03374.

[13] Rishi Bommasani, Drew A. Hudson, and Ehsan Adeli, "On the Opportunities and Risks of Foundation Models," arXiv (2021), https://arxiv.org/abs/2108.07258v1.

Searching for Computer Vision North Stars

Li Fei-Fei & Ranjay Krishna

Computer vision is one of the most fundamental areas of artificial intelligence research. It has contributed to the tremendous progress in the recent deep learning revolution in AI. In this essay, we provide a perspective of the recent evolution of object recognition in computer vision, a flagship research topic that led to the breakthrough data set of ImageNet and its ensuing algorithm developments. We argue that much of this progress is rooted in the pursuit of research "north stars," wherein researchers focus on critical problems of a scientific discipline that can galvanize major efforts and groundbreaking progress. Following the success of ImageNet and object recognition, we observe a number of exciting areas of research and a growing list of north star problems to tackle. This essay recounts the brief history of ImageNet, its related work, and the follow-up progress. The goal is to inspire more north star work to advance the field, and AI at large.

Artificial intelligence is a rapidly progressing field. To many of its everyday users, AI is an impressive feat of engineering derived from modern computer science. There is no question that there has been incredible engineering progress in AI, especially in recent years. Successful implementations of AI are all around us, from email spam filters and personalized retail recommendations to cars that avoid collisions in an emergency by autonomously braking. What may be less obvious is the science behind the engineering. As researchers in the field, we have a deep appreciation of both the engineering and the science and see the two approaches as deeply intertwined and complementary. Thinking of AI, at least in part, as a scientific discipline can inspire new lines of thought and inquiry that, in time, will make engineering progress more likely. As in any science, it is not always obvious what problems in AI are the most important to tackle. But once you have formulated a fundamental problem – once you have identified the next "north star" – you can start pushing the frontier of your field. That has certainly been our experience, and it is why we love Einstein's remark that "The mere formulation of a problem is often far more essential than its solution."

AI has been driven by north stars from the field's inception in 1950, when Alan Turing neatly formulated the problem of how to tell if a computer deserves to be called intelligent. (The computer, according to the now-famous Turing Test,

© 2022 by Li Fei-Fei & Ranjay Krishna
Published under a Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) license
https://doi.org/10.1162/DAED_a_01902

would need to be able to "deceive a human into believing that it was human," as Turing put it.)[1] A few years later, as the founding fathers of AI planned the Dartmouth workshop, they set another ambitious goal, proposing to build machines that can "use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves."[2] Without that guiding light, we might never be in a position to tackle new problems.

Our own area within AI, computer vision, has been driven by its own series of north stars. This is the story of one – object recognition – and the progress it has made toward north stars in other AI fields.

The ability to see – vision – is central to intelligence. Some evolutionary biologists have hypothesized that it was the evolution of eyes in animals that first gave rise to the many different species we know today, including humans.[3]

Seeing is an immensely rich experience. When we open our eyes, the entire visual world is immediately available to us in all its complexity. From registering shadows and brightness, to taking in the colors of everything around us, to recognizing an appetizing banana on a kitchen counter as something good to eat, humans use our visual perception to navigate the world, to make sense of it, and to interact with it. So how do you even begin to teach a computer to see? There are many important problems to solve and choosing them is an essential part of the scientific quest for computer vision: that is, the process of identifying the north stars of the field. At the turn of the century, inspired by a large body of important work prior to ours, our collaborators and we were drawn to the problem of object recognition: a computer's ability to correctly identify what appears in a given image.

This seemed like the most promising north star for two reasons. The first was its practical applications. The early 2000s witnessed an explosive increase in the number of digital images, thanks to the extraordinary growth of the Internet and digital cameras, and all those images created a demand for tools to automatically catalog personal photo collections and to enable users to search through such image collections. Both applications would require object recognition.

But an even deeper reason was the remarkable ability of humans to perceive and interpret objects in the visual world. Research in the field of cognitive neuroscience showed that humans can detect animals within just twenty milliseconds and, within only three hundred milliseconds, can tell whether the animal is, say, a tiger or a lamb. The research in cognitive neuroscience also offered clues to how humans are able to achieve such rapid recognition: scientists had found that humans relied on cues in the object's surroundings and on certain key features of objects, features that did not change with a difference in angle or lighting conditions. Most strikingly, neuroscientists had discovered specific regions of the brain that activate when people view specific objects.[4] The existence of neural correlates for

any function is a sure sign of the function's evolutionary importance: a specific brain region would not evolve for a specific function unless that function was essential for the organism's survival or reproduction. Clearly, the ability to recognize specific objects must be critical.

These findings made clear to us that object recognition should be considered a north star in computer vision. But how do you get a computer to recognize objects? Recognizing objects requires understanding what concept a digital image represents in the visual world – what the image *means* – but a computer has no such understanding. To a computer, a digital image is nothing more than a collection of pixels, a two-dimensional array of numbers that does not really mean anything except colors and illuminations. Teaching a computer to recognize objects requires somehow getting it to connect each lifeless collection of numbers to a meaningful concept, like dog or banana.

Between the decades of the 1990s and the early 2000s, researchers in object recognition had already made tremendous progress toward this daunting goal, but progress was slow because of the enormous variety in the appearance of real-world objects. Even within a single, fairly specific category (like house, dog, or flower), objects can look quite different. For example, an AI capable of accurately recognizing an object in a photograph as a dog needs to recognize it as a dog whether it is a German shepherd, poodle, or chihuahua. And whatever the breed, the AI needs to recognize it as a dog whether it is photographed from the front or from the side, running to catch a ball or standing on all fours with a blue bandana around its neck. In short, there is a bewildering diversity of images of dogs, and past attempts at teaching computers to recognize such objects failed to cope with this diversity.

One major bottleneck of most of these past methods was their reliance on hand-designed templates to capture the essential features of an object, and the lack of exposure to a vast variety of images. Computers learn from being exposed to examples; that is the essence of machine learning. And while humans can often generalize correctly from just a few examples, computers need large numbers of examples; otherwise, they make mistakes. So AI researchers had been trapped in a dilemma. On the one hand, for a template to be helpful in teaching an AI system to recognize objects, the template needed to be based upon a large variety of images and, therefore, a very large number of images in total. On the other hand, hand-designing a template is labor-intensive work, and doing so from a very large number of images is not feasible.

The inability to scale the template approach effectively made it clear that we needed a different way to approach the object-recognition problem.

We started our search for a new approach with one key assumption: even the best algorithm would not generalize well if the data it learned from did not reflect the real world. In concrete terms, that meant that ma-

jor advances in object recognition could occur only from access to a large quantity of diverse, high-quality training data. That assumption may sound obvious because we are all awash in data and we all benefit from powerful object-recognition tools. But when we began our work in the early 2000s, the focus on data was fairly contrarian: at that time, most people in our field were paying attention to models (algorithms), not to data. Of course, in truth, the two pursuits are compatible. We believed that good data would help with the design of good models, which would lead to advances in object recognition and in AI more broadly.

That meant that we needed to create a new data set (which we called ImageNet) that achieved these three design goals: scale (a large quantity of data), diversity (a rich variety of objects), and quality (accurately labeled objects).[5] In focusing on these three goals, we had moved from a general north star – image recognition – to more specific problem formulations. But how did we tackle each?

Scale. Psychologists have posited that human-like perception requires exposure to thousands of diverse objects.[6] When young children learn naturally, their lives have already been exposed to enormous numbers of images every day. For example, by the time a typical child is six years old, she has seen approximately three thousand distinct objects, according to one estimate; from those examples, the child would have learned enough distinctive features to help distinguish among thirty thousand more categories. That is how large a scale we had in mind. Yet the most popular object-recognition data set when we began included only twenty objects, the result of the very process we described earlier as too cumbersome to scale up. Knowing that we needed far more objects, we collected fifteen million images from the Internet.

But images alone would not be enough to provide useful training data to a computer: we would also need meaningful categories for labeling the objects in these images. After all, how can a computer know that a picture of a dog is a German shepherd (or even a dog) unless the picture has been labeled with one of these categories? Furthermore, most of the machine learning algorithms require a training phase during which the algorithms must learn from labeled examples (that is, training examples) and be measured by their performances of a separate set of labeled examples (that is, testing samples). So we turned to an English-language vocabulary data set, called WordNet, developed by cognitive psychologist George Miller in 1990.[7] WordNet organizes words into hierarchically nested categories (such as dog, mammal, and animal); using WordNet, we chose thousands of object categories that would encompass all the images we had found. In fact, we named our data set ImageNet by analogy with WordNet.

Diversity. The images we collected from the Internet represented the diversity in real-world objects, covering many categories. For example, there were more than eight hundred different kinds of birds alone, with several examples of each. In total, we used 21,841 categories to organize the fifteen million images in our

data set. The challenges in capturing real-world diversity within each category is that simple Internet search results are biased toward certain kinds of images: for example, Google's top search results for "German shepherd" or "poodle" consist of cleanly centered images of each breed. To avoid this kind of bias, we had to expand the query to include a description: to search also, for example, for "German shepherd in the kitchen." Similarly, to get a broader, more representative distribution of the variety of dog images, we used translations into some other languages as well as hypernyms and hyponyms: not just "husky" but also "Alaskan husky" and "heavy-coated Arctic sled dog."

Quality. We cared a lot about the quality of the images and the quality of the annotations. To create a gold-standard data set that would replicate the acuity of human vision, we used only high-resolution images. And to create accurate labels for the objects in the data set, we hired people. At first, we brought in Princeton undergraduate students to label the images and verify these labels, but it quickly became apparent that using such a small group would take far too long. Through a fortunate coincidence, Amazon had just released its crowdsourcing platform, Mechanical Turk, which enabled us to quickly hire approximately fifty thousand workers from 167 countries to label and verify the objects in our set between 2007 and 2009.[8]

The ImageNet team believed it was important to democratize research in object recognition and to build a community around ImageNet. So we open-sourced ImageNet: we made it free and open to any interested researcher. We also established an annual competition to inspire researchers from all around the world. The ImageNet Large-Scale Visual Recognition Challenge (often simply called the ImageNet Challenge), which ran concurrently from 2010 until 2017 with the international computer vision research conferences International Conference on Computer Vision and European Conference on Computer Vision, created a common benchmark for measuring progress.

We set up the ImageNet Challenge similar to the design of other machine learning competitions: All participants would get the same training data, which is just a subset of the larger ImageNet data set. After using this training data to train their object-recognition algorithm, the participants would unleash their algorithm on unlabeled images that the algorithm had never encountered to see how accurately the algorithm would recognize these new images. These test data, too, came from ImageNet.

We had high aspirations for the ImageNet data set and for the ImageNet Challenge, yet the outcomes exceeded them. The biggest turning point came in 2012, when one team applied a convolutional neural network to object recognition for the first time.[9] (A convolutional neural network is an algorithm inspired by the way the human brain works.) That team's winning entry, later known as AlexNet

after one of its creators, trounced its competition, recognizing images with an accuracy that was a whopping 41 percent higher than that of the second-place finisher. Although neural networks as an approach to machine learning had been around for decades, it had not been widely used until that year's ImageNet Challenge.

This was a watershed moment for the AI community. The impressive performance of AlexNet on the ImageNet data set inspired other researchers – and not just participants in the ImageNet Challenge – to shift to deep learning approaches. We started seeing large companies like Google and Facebook deploying technology based on neural networks, and within a year, almost every AI paper was about neural networks.

With so many people working on neural networks, the technology advanced rapidly. Researchers found that the deeper the model, the better it performed at object recognition. And as deeper models required more processing power, researchers ran into other problems, such as computational bottlenecks, which required further design work to overcome. The ImageNet Challenge created a kind of domino effect of innovations, with each advance leading to more.[10]

Beyond the tremendous progress in computer vision through more and more powerful deep learning algorithms, researchers began using deep learning to automate and systematize the design of model architecture itself, instead of hand-designing each neural network's architecture. The process of hand-designing architectures, like the previous process of hand-designing features in templates, is speculative: the search space of possible architectures is exponentially vast, so manual architectural changes are unlikely to thoroughly explore this space quickly enough to uncover the optimal architecture. Using ImageNet as a test bed, computer vision researchers have systematized the process of neural architecture search.[11] Initial methods consumed too many computational resources to exhaustively cover the search space. Inspired by the success of hand-designed architectures with recurring architecture motifs, such as ResNet36 and Inception35, later methods defined architectures with recurring cell structures and restricted the search space to designing this recurring cell.[12]

The ImageNet Challenge ended once the accuracy of its best models reached superhuman levels, at 97.3 percent. (Human accuracy on this data was about 95 percent.)[13] Other researchers have continued making incremental advancements, however, using the ImageNet data set to track their progress, and error rates have continued to fall, though certainly not as fast as in the first few years after the introduction of ImageNet. The error rate of the best model today is only 1.2 percent, down from 33.6 percent when the competition began back in 2009.[14]

These days, thanks to high accuracy and reasonable computing costs, object recognition is in wide use. Whenever you search for images on the Internet, you use the kinds of algorithms first developed for the ImageNet Challenge; the same goes for when your smartphone automatically groups your photos based on

whose face appears in the photo. Those are exactly the uses we had in mind when we first chose object recognition as our north star. But uses of object recognition go beyond that, from tracking players in sports to helping self-driving cars detect other vehicles.

Learning to recognize objects is only one form of learning to see, which is why computer vision (or visual intelligence) is a much broader field than object recognition. But there are important similarities between object recognition and other tasks in computer vision, such as object detection and activity recognition. Such similarities mean that a computer should not need to tackle a new task from scratch. In theory, a computer should be able to take advantage of the similarities, applying what it has learned from one task to perform a somewhat different task. For both computers and humans, this process of generalizing knowledge from one task to a similar one is called *transfer learning*.[15]

Humans are very good at transfer learning: once we know French, for example, it is not as hard to learn Spanish. And if you learned to read English as a child, that was certainly easier if you already knew how to speak English than if the language was entirely new to you. In fact, the ability to pick up on similarities between tasks, and to parlay this shared knowledge to help us learn new tasks, is one of the hallmarks of human intelligence.

Transfer learning can be tremendously helpful for AI, too, but it does not come naturally to computers; instead, we humans have to teach them. The way to help computers with transfer learning is through pretraining. The idea is that before you give a machine learning model a new challenge, you first train it to do something similar, using training data that are already known to be effective. In computer vision, that starting point is the object-recognition data in ImageNet. Once a new model gets trained through ImageNet, it should have a leg up on tackling a new kind of challenge. If this approach works, as we thought it would, then we have all the more reason to think that object recognition is a north star for visual intelligence.

That was the thinking behind our extension of the ImageNet Challenge to the problem of object detection. Object detection means recognizing an object in an image and specifying its location within the image. If you have ever seen a digital photograph of a group of people with a little rectangle drawn around each person's face, you have seen one application of object detection. Whereas the images in ImageNet contain just one object each, most real-world scenes include several objects, so object detection is a valuable extension of the kind of simple object recognition we had tested in the ImageNet Challenge.

Object detection had been an area of research before ImageNet, too, but the most common approach then was to first identify the areas within the image where an object (such as an animal) was likely to be, and then to focus on that area

and try to recognize that object (as a tiger, for example).[16] Once ImageNet became available, that second step became much easier.

Object detection has come a long way since then, with special-purpose detectors for different kinds of applications, such as self-driving cars, which need to be alert to other cars on the road.[17] Such advances beyond object recognition would not have been possible without the use of ImageNet to enable transfer learning.

But object detection was just a first attempt to apply ImageNet data to uses beyond object recognition. These days, for better or for worse, almost every computer vision method uses models pretrained on ImageNet.

None of that is to say that ImageNet has been useful for every computer vision task. A prominent example is medical imaging.[18] Conceptually, the task of classifying a medical image (such as a screening mammogram) is not very different from the task of classifying a photograph taken with a phone camera (such as a snapshot of a family pet). Both tasks involve visual objects and category labels, so both could be performed by a properly trained machine. In fact, they have been. But the methods have not been exactly the same. For one thing, you cannot use the ImageNet data set to train a computer to detect tumors; it simply has no data for this specialized task. What is more, it is not feasible to use the same basic approach: the professional expertise required to create high-quality training data to help with medical diagnosis is scarce and expensive. Put another way, it is impossible to use Mechanical Turk to create a high-quality medical data set, both due to the requirement of specialized expertise as well as regulatory restrictions. So instead of using carefully labeled examples (the process of "supervised learning"), AI for medical imaging is usually based on "semi-supervised learning," whereby the machine learns to find meaningful patterns across images without many explicit labels.[19]

Computer vision certainly has practical applications beyond health, including environmental sustainability. Researchers are already using machine learning to analyze large volumes of satellite images to help governments assess changes in crop yields, water levels, deforestation, and wildfires, and to track longitudinal climate change.[20] Computer vision can be helpful in education, too: when students are trying to learn to read bar charts or to study visual subjects like geometry and physics, computers that understand images have the potential to supplement the efforts of human teachers. Assistive technology could also help teachers generate content-appropriate quizzes.[21]

The use of ImageNet to generalize beyond object recognition also led to the discovery of a thorny problem for deep learning models: "adversarial examples," which are images that fool an AI into making blatant errors classifying an object.[22] A miniscule, humanly imperceptible tweak to a picture (sometimes even a single pixel!) can cause a model trained on ImageNet to mislabel it entirely.[23] An image

of a panda can thus get misclassified as a bathtub. Some kinds of errors are easier to understand as the result of spurious correlations: wolves are often photographed in snow, so a model that learns to associate snow with wolves could come to assume that the label "wolf" refers to "snow." It turns out that all models that use deep learning are vulnerable to attacks from adversarial examples, a fact that has spurred some researchers to work on ways to "vaccinate" training data against these attacks.

The problem of adversarial examples has also led the computer vision community to shift from a singular focus on accuracy. Although accuracy in object recognition certainly remains important, researchers have come to appreciate the value of other criteria for evaluating a machine learning model, particularly interpretability (which refers to the ability of a model to generate predictable or understandable inference results for human beings) and explainability (the ability of a model to provide post hoc explanations for existing black box models).[24]

The success of ImageNet has also prompted the computer vision community to start asking what data the next generation of models should be pretrained on. As an alternative to the expensive, carefully annotated, and thoroughly verified process used to create ImageNet, researchers have collected data from social media and scraped images with their associated text off the Internet.[25] Pretraining models from this "raw" data have opened up the possibility of "zero-shot adaptation," the process through which computers can learn without any explicit labels. In fact, models trained on such raw data now perform as well as models trained using ImageNet.[26]

Finally, the wide influence of ImageNet has opened the data set up to criticism, raising valid concerns we were not sufficiently attuned to when we began. The most serious of these is the issue of fairness in images of people.[27] For one thing, although we certainly knew early on to filter out blatantly derogatory image labels such as racial or gender slurs, we were not sensitive to more subtle problems, such as labels that are not inherently derogatory but could cause offense when applied inappropriately (such as labeling people based on clues to their religion or sexual orientation). In addition, certain concepts related to people are hard to represent visually without resorting to stereotypes, so attempts to associate images with these concept labels ("philanthropist" or "Bahamian," for example) perpetuate biases. Most Bahamian wear distinctive garb only on special, ceremonial occasions, but an image search for "Bahamian" based on ImageNet data would give a disproportionate number of such stereotypical images of people from the Bahamas. Another source of bias in search results is the inadequate diversity in the ImageNet data set, a bias that tends to get amplified during the manual cleanup stage, when human annotators resort to racial and gender stereotypes in their labeling. Women and ethnic minorities are already underrepresented among real-world bankers, for example, but they are even *more* underrepresented in images

labeled as "banker." Although these problems of fairness are difficult to eliminate entirely, we have made research strides to mitigate them.[28]

The development of these new data sets has led to the need for a metabenchmark: a single evaluation scheme for multiple individual benchmarks (or a benchmark for comparing benchmarks). Without a metabenchmark, it is impossible to compare the performance of different machine learning models across different tasks and using different data sets.

In fact, one thing that has emerged is a lively debate about benchmarks themselves.[29] One side of the debate posits that the constant emergence of new benchmarks is a good sign, suggesting continued progress on north stars. On the other side is a concern that benchmarks encourage something akin to teaching to the test: the concern that what emerges from benchmarking are not superior models but models that optimize for high performance on an inherently imperfect benchmark.

Another serious concern is that a widely adopted benchmark amplifies the real-world effects of any flaws in the benchmark. There is a growing body of research, for example, on how benchmarks can perpetuate structural societal biases,[30] benefiting groups that are already dominant (particularly White males) while discriminating against marginalized groups (such as Muslims and dark-skinned females).[31]

In response to these concerns, pioneers in the field are radically rethinking benchmarking. One suggestion has been for human judges to generate inputs for which models would fail, thus creating increasingly harder testing criteria as models improve.[32] Another idea is to demand that benchmarks measure not only accuracy (which encourages designing to the benchmark) but also assess and reward progress on other valuable criteria, including bias detection.[33]

Where do we go next in computer vision? Other north stars beckon. One of the biggest is in the area of embodied AI: robotics for tasks such as navigation, manipulation, and instruction following. That does not necessarily mean creating humanoid robots that nod their heads and walk on two legs; any tangible and intelligent machine that moves through space is a form of embodied AI, whether it is a self-driving car, a robot vacuum, or a robotic arm in the factory. And just as ImageNet aimed at representing a broad and diverse range of real-world images, research in embodied AI needs to tackle the complex diversity of human tasks, from folding laundry to exploring a new city.[34]

Another north star is visual reasoning: understanding, for example, the three-dimensional relationships in a two-dimensional scene. Think of the visual reasoning needed to follow even the seemingly simple instruction to bring back the metal mug to the left of the cereal bowl. Following such instructions certainly requires more than vision, but vision is an essential component.[35]

Understanding people in a scene, including social relationships and human intentions, adds yet another level of complexity, and such basic social intelligence is another north star in computer vision.[36] Even a five-year-old can guess, for example, that if a woman is cuddling with a little girl on her lap, the two people are very likely mother and daughter, and that if a man opens a refrigerator, he is probably hungry; but computers do not yet have enough intelligence to infer such things. Computer vision, like human vision, is not just perception; it is deeply cognitive.

There is no question that all these north stars are huge challenges, bigger than ImageNet ever was. It is one thing to review photos to try to identify dogs or chairs, and it is another to think about and navigate the infinite world of people and space. But it is a set of challenges well worth pursuing: as computers' visual intelligence unfolds, the world can become a better place. Doctors and nurses will have extra pairs of tireless eyes to help them diagnose and treat patients. Cars will run more safely. Robots will help humans brave disaster zones to save the trapped and wounded. And scientists, with help from machines that can see what humans cannot, will discover new species, better materials, and uncharted frontiers.

AUTHORS' NOTE

The authors are grateful for the support of the Office of Naval Research MURI grants, a Brown Institute grant, the Stanford Institute for Human-Centered Artificial Intelligence, and the Toyota Research Institute.

ABOUT THE AUTHORS

Li Fei-Fei, a Fellow of the American Academy since 2021, is the Sequoia Capital Professor and Denning Family Co-Director of the Stanford Institute for Human-Centered Artificial Intelligence. She is an elected member of the National Academy of Engineering and the National Academy of Medicine. She has published in such journals as *Nature*, *Proceedings of the National Academy of Sciences*, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *International Journal of Robotics Research*, *International Journal of Computer Vision*, and *The New England Journal of Medicine*.

Ranjay Krishna is an Assistant Professor at the Allen School of Computer Science & Engineering at the University of Washington. He has published essays in such journals as *International Journal of Computer Vision* and academic book chapters with Springer Science-Business Media, and has presented academic conference papers at top-tier computing venues for computer vision, natural language processing, and human-computer interaction.

ENDNOTES

[1] Alan M. Turing, "Computing Machinery and Intelligence," *Mind* 59 (236) (1950): 433–460.

[2] John McCarthy, Marvin L. Minsky, Nathaniel Rochester, and Claude E. Shannon, "A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence," *AI Magazine*, Winter 2006.

[3] Andrew Parker, *In the Blink of an Eye: How Vision Started the Big Bang of Evolution* (London: Simon & Schuster, 2003), 316.

[4] Nancy Kanwisher, Josh McDermott, and Marvin M. Chun, "The Fusiform Face Area: A Module in Human Extrastriate Cortex Specialized for Face Perception," *Journal of Neuroscience* 17 (11) (1997): 4302–4311; and Russell Epstein and Nancy Kanwisher, "A Cortical Representation of the Local Visual Environment," *Nature* 392 (6676) (1998): 598–601.

[5] Jia Deng, Wei Dong, Richard Socher, et al., "ImageNet: A Large-Scale Hierarchical Image Database," in 2009 *IEEE Conference on Computer Vision and Pattern Recognition* (Red Hook, N.Y.: Curran Associates, Inc., 2009), https://image-net.org/static_files/papers/imagenet_cvpr09.pdf.

[6] Irving Biederman, "Recognition-by-Components: A Theory of Human Image Understanding," *Psychological Review* 94 (2) (1987), https://psycnet.apa.org/record/1987-20898-001.

[7] George A. Miller, "WordNet: A Lexical Database for English," *Communications of the ACM* 38 (11) (1995): 39–41.

[8] Jia Deng, Olga Russakovsky, Jonathan Krause, et al., "Scalable Multi-Label Annotation," in *CHI '14: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (New York: Association for Computing Machinery, 2014), 3099–3102.

[9] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *Advances in Neural Information Processing Systems* (2012): 1097–1105.

[10] Ibid.; Christian Szegedy, Wei Liu, Yangqing Jia, et al., "Going Deeper with Convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Piscataway, N.J.: Institute of Electrical and Electronics Engineers, 2015), 1–9; Karen Simonyan and Andrew Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," presented at the 2015 International Conference on Learning Representations, San Diego, California, May 7, 2015; Kaiming He, Xiangyu Zhang, Shaoqing Ren, et al., "Deep Residual Learning for Image Recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Piscataway, N.J.: Institute of Electrical and Electronics Engineers, 2016); Saining Xie, Ross Girshick, Piotr Dollár, et al., "Aggregated Residual Transformations for Deep Neural Networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Piscataway, N.J.: Institute of Electrical and Electronics Engineers, 2017); and Chenxi Liu, Barret Zoph, Maxim Neumann, et al., "Progressive Neural Architecture Search," in *Proceedings of the European Conference on Computer Vision* (Cham, Switzerland: Springer Nature, 2018).

[11] Barret Zoph and Quoc V. Le, "Neural Architecture Search with Reinforcement Learning," presented at the 5th International Conference on Learning Representations, Toulon, France, April 26, 2017; and Hieu Pham, Melody Guan, Barret Zoph, et al., "Efficient Neural Architecture Search via Parameters Sharing," *Proceedings of Machine Learning Research* 80 (2018): 4095–4104.

[12] Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc V. Le, "Learning Transferable Architectures for Scalable Image Recognition," in *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Piscataway, N.J.: Institute of Electrical and Electronics Engineers, 2018), 8697–8710; and Hanxiao Liu, Karen Simonyan, Oriol Vinyals, et al., "Hierarchical Representations for Efficient Architecture Search," arXiv (2017), https://arxiv.org/abs/1711.00436.

[13] Olga Russakovsky, Jia Deng, Hao Su, et al., "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision* 115 (3) (2015): 211–252.

[14] Hieu Pham, Zihang Dai, Qizhe Xie, et al., "Meta Pseudo Labels," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Piscataway, N.J.: Institute of Electrical and Electronics Engineers, 2021), 11557–11568.

[15] Yusuf Aytar and Andrew Zisserman, "Tabula Rasa: Model Transfer for Object Category Detection," presented at the 2011 International Conference on Computer Vision, Barcelona, Spain, November 6–13, 2011; Maxime Oquab, Leon Bottou, Ivan Laptev, et al., "Learning and Transferring Mid-Level Image Representations Using Convolutional Neural Networks," in *2014 IEEE Conference on Computer Vision and Pattern Recognition* (Piscataway, N.J.: Institute of Electrical and Electronics Engineers, 2014), 1717–1724; and Zhizhong Li and Derek Hoiem, "Learning without Forgetting," in *Proceedings of the European Conference on Computer Vision* (Cham, Switzerland: Springer Nature, 2016).

[16] Shaoqing Ren, Kaiming He, Ross Girshick, et al., "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," *Advances in Neural Information Processing Systems* 28 (2015): 91–99; and David A. Forsyth and Jean Ponce, *Computer Vision: A Modern Approach* (Hoboken, N.J.: Prentice Hall, 2011).

[17] Bichen Wu, Forrest Iandola, Peter H. Jin, and Kurt Keutzer, "SqueezeDet: Unified, Small, Low Power Fully Convolutional Neural Networks for Real-Time Object Detection for Autonomous Driving," in *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops* (Piscataway, N.J.: Institute of Electrical and Electronics Engineers, 2017), 129–137.

[18] Varun Gulshan, Lily Peng, Marc Coram, et al., "Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs," *The Journal of the American Medical Association* 316 (22) (2016): 2402–2410; Geert Litjens, Thijs Kooi, and Babak Ehteshami Bejnordi, "A Survey on Deep Learning in Medical Image Analysis," *Medical Image Analysis* 42 (2017): 60–88; and Hayit Greenspan, Bram van Ginneken, and Ronald M. Summers, "Guest Editorial: Deep Learning in Medical Imaging: Overview and Future Promise of an Exciting New Technique," *IEEE Transactions on Medical Imaging* 35 (5) (2016): 1153–1159.

[19] Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien, "Semi-Supervised Learning," *IEEE Transactions on Neural Networks* 20 (3) (2009): 542–542; David Berthelot, Nicholas Carlini, Ian Goodfellow, et al., "Mixmatch: A Holistic Approach to Semi-Supervised Learning," *Advances in Neural Information Processing Systems* 32 (2019); and Kihyuk Sohn, David Berthelot, Chun-Liang Li, et al., "FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence," *Advances in Neural Information Processing Systems* 33 (2020).

[20] Neal Jean, Marshall Burke, Michael Xie, et al., "Combining Satellite Imagery and Machine Learning to Predict Poverty," *Science* 353 (6301) (2016): 790–794.

[21] Chris Piech, Jonathan Spencer, Jonathan Huang, et al., "Deep Knowledge Tracing," *Advances in Neural Information Processing Systems* 28 (2015).

[22] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, et al., "Intriguing Properties of Neural Networks," arXiv (2013), https://arxiv.org/abs/1312.6199.

[23] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy, "Explaining and Harnessing Adversarial Examples," presented at the 2015 International Conference on Learning Representations, San Diego, California, May 9, 2015; Szegedy et al., "Intriguing Properties of Neural Networks"; and Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, et al., "Towards Deep Learning Models Resistant to Adversarial Attacks," presented at the Sixth International Conference on Learning Representations, Vancouver, Canada, April 30, 2018.

[24] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin, "'Why Should I Trust You?' Explaining the Predictions of Any Classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (New York: Association for Computing Machinery, 2016); Cynthia Rudin, "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead," *Nature Machine Intelligence* 1 (5) (2019): 206–215; and Ričards Marcinkevičs and Julia E. Vogt, "Interpretability and Explainability: A Machine Learning Zoo Mini-Tour," arXiv (2020), https://arxiv.org/abs/2012.01805.

[25] Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, et al., "Exploring the Limits of Weakly Supervised Pretraining," in *Proceedings of the 2018 European Conference on Computer Vision* (New York: Computer Vision Foundation, 2018), 181–196; Ranjay Krishna, Yuke Zhu, Oliver Groth, et al., "Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations," *International Journal of Computer Vision* 123 (2017): 32–73; and Tsung-Yi Lin, Michael Maire, Serge Belongie, et al., "Microsoft COCO: Common Objects in Context," in *Proceedings of the European Conference on Computer Vision* (Cham, Switzerland: Springer Nature, 2014).

[26] Alec Radford, Jong Wook Kim, Chris Hallacy, et al., "Learning Transferable Visual Models from Natural Language Supervision," arXiv (2021), https://arxiv.org/abs/2103.00020; Zirui Wang, Jiahui Yu, Adams Wei Yu, et al., "Simvlm: Simple Visual Language Model Pretraining with Weak Supervision," arXiv (2021); and Chao Jia, Yinfei Yang, Ye Xia, et al., "Scaling Up Visual and Vision-Language Representation Learning with Noisy Text Supervision," arXiv (2021), https://arxiv.org/abs/2102.05918.

[27] Chris Dulhanty and Alexander Wong, "Auditing ImageNet: Towards a Model-Driven Framework for Annotating Demographic Attributes of Large-Scale Image Datasets," arXiv (2019), https://arxiv.org/abs/1905.01347; Pierre Stock and Moustapha Cisse, "ConvNets and ImageNet Beyond Accuracy: Understanding Mistakes and Uncovering Biases," in *Proceedings of the European Conference on Computer Vision* (Cham, Switzerland: Springer Nature, 2018); and Eran Eidinger, Roee Enbar, and Tal Hassner, "Age and Gender Estimation of Unfiltered Faces," *IEEE Transactions on Information Forensics and Security* 9 (12) (2014): 2170–2179.

[28] Kaiyu Yang, Klint Qinami, Li Fei-Fei, et al., "Towards Fairer Datasets: Filtering and Balancing the Distribution of the People Subtree in the ImageNet Hierarchy," in *FAT '20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (New York: Association for Computing Machinery, 2020), 547–558.

[29] Sanjeev Arora and Yi Zhang, "Rip van Winkle's Razor: A Simple Estimate of Overfit to Test Data," arXiv (2021), https://arxiv.org/abs/2102.13189; and Avrim Blum and Moritz Hardt, "The Ladder: A Reliable Leaderboard for Machine Learning Competitions," *Proceedings of Machine Learning Research* 37 (2015): 1006–1014.

[30] Antonio Torralba and Alexei A. Efros, "Unbiased Look at Dataset Bias," in *Proceedings of the 2011 Conference on Computer Vision and Pattern Recognition* (Piscataway, N.J.: Institute of Electrical and Electronics Engineers, 2011), 1521–1528; Vinay Uday Prabhu and Abeba Birhane, "Large Image Datasets: A Pyrrhic Win for Computer Vision?" arXiv (2020), https://arxiv.org/abs/2006.16923; and Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan, "Semantics Derived Automatically from Language Corpora Contain Human-Like Biases," *Science* 356 (6334) (2017): 183–186, https:// science.sciencemag.org/content/356/6334/183.

[31] Abubakar Abid, Maheen Farooqi, and James Zou, "Persistent Anti-Muslim Bias in Large Language Models," arXiv (2021), https://arxiv.org/abs/2101.05783.

[32] Ozan Sener and Silvio Savarese, "Active Learning for Convolutional Neural Networks: A Core-Set Approach," presented at the Sixth International Conference on Learning Representations, Vancouver, Canada, May 1, 2018; and Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick, "Training Region-Based Object Detectors with Online Hard Example Mining," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Piscataway, N.J.: Institute of Electrical and Electronics Engineers, 2016).

[33] Moin Nadeem, Anna Bethke, and Siva Reddy, "Stereoset: Measuring Stereotypical Bias in Pretrained Language Models," arXiv (2020), https://arxiv.org/abs/2004.09456; and Timnit Gebru, "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification," *Proceedings of Machine Learning Research* 81 (2018): 77–91.

[34] Sanjana Srivastava, Chengshu Li, Michael Lingelbach, et al., "BEHAVIOR: Benchmark for Everyday Household Activities in Virtual, Interactive, and Ecological Environments," presented at the Conference on Robot Learning, London, England, November 9, 2021; Eric Kolve, Roozbeh Mottaghi, Winson Han, et al., "AI2-THOR: An Interactive 3D Environment for Visual AI," arXiv (2017), https://arxiv.org/abs/1712.05474; and Xavier Puig, Kevin Ra, Marko Boben, et al., "VirtualHome: Simulating Household Activities via Programs," in *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Piscataway, N.J.: Institute of Electrical and Electronics Engineers, 2018).

[35] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, et al., "CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Piscataway, N.J.: Institute of Electrical and Electronics Engineers, 2017); Adam Santoro, David Raposo, David G. T. Barrett, et al., "A Simple Neural Network Module for Relational Reasoning," presented at the 31st Conference on Neural Information Processing Systems, Long Beach, California, December 6, 2017; and Justin Johnson, Bharath Hariharan, Laurens van der Maaten, et al., "Inferring and Executing Programs for Visual Reasoning," in *Proceedings of the 2017 IEEE International Conference on Computer Vision* (Piscataway, N.J.: Institute of Electrical and Electronics Engineers, 2017).

[36] Alexandre Alahi, Kratarth Goel, Vignesh Ramanathan, et al., "Social LSTM: Human Trajectory Prediction in Crowded Spaces," in *2016 IEEE Conference on Computer Vision and Pattern Recognition* (Piscataway, N.J.: Institute of Electrical and Electronics Engineers, 2016); and Alina Kuznetsova, Hassan Rom, Neil Alldrin, et al., "The Open Images Dataset V4: Unified Image Classification, Object Detection, and Visual Relationship Detection at Scale," *International Journal of Computer Vision* 128 (7) (2020).

The Machines from Our Future

Daniela Rus

While the last sixty years have defined the field of industrial robots and empowered hard-bodied robots to execute complex assembly tasks in constrained industrial settings, the next sixty years will usher in our time with pervasive robots that come in a diversity of forms and materials and help people with physical tasks. The past sixty years have mostly been inspired by the human form, but the form diversity of the animal kingdom has broader potential. With the development of soft materials, machines and materials are coming closer together: machines are becoming compliant and fluid-like materials, and materials are becoming more intelligent. This progression raises the question: what will be the machines from our future?

Today, telepresence enables students to meet with tutors and teachers and allows doctors to treat patients thousands of miles away. Robots help with packing on factory floors. Networked sensors enable the monitoring of facilities, and 3D printing creates customized goods. We are surrounded by a world of possibilities. And these possibilities will only get larger as we start to imagine what we can do with advances in artificial intelligence and robotics. Picture a world where routine tasks are taken off your plate. Fresh produce just shows up on your doorstep, delivered by drones. Garbage bins take themselves out, and smart infrastructure systems support automated pick-up. AI assistants – whether embodied or not – act as guardian angels, providing advice to ensure that we maximize and optimize our lives to live well and work effectively.

The field of robotics has the potential to greatly improve the quality of our lives at work, at home, and at play by providing people with support for cognitive and physical tasks. For years, robots have supported human activity in dangerous, dirty, and dull tasks, and have enabled the exploration of unreachable environments, from the deep oceans to deep space. Increasingly more-capable robots will be able to adapt, learn, and interact with humans and other machines on cognitive levels. The objective of robotics is not to replace humans by mechanizing and automating tasks, but rather to find new ways that allow robots to collaborate with humans more effectively. Machines are better than humans at tasks such as crunching numbers and moving with precision. Robots can lift much heavier objects. Humans are better than machines at tasks like reasoning, defining abstractions, and generalizing or specializing, thanks to our ability to draw on prior expe-

© 2022 by Daniela Rus
Published under a Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) license
https://doi.org/10.1162/DAED_a_01903

riences. By working together, robots and humans can augment and complement each other's skills.

Imagine riding in your flying car, which is integrated with the information technology infrastructure and knows your needs, so it can tell you, for example, that you can buy the plants you have been wanting at a store nearby, while computing a small detour. You can trust your home to take care of itself when you are away. That is what the smart refrigerator is for: it tracks everything you put in and take out so it can automatically send a shopping list to your favorite delivery service when it is time to restock. This automated household can help take care of everything from your new plants to your elderly parents. The intelligent watering system monitors the soil and ensures each type of plant gets the right level of moisture. When your elderly parents need help with cooking, the kitchen robot can assist. The new robotic technologies can also be carried with us, knitted in our sweaters, blended in our garments, or embedded in our accessories. We could have our own wearable computer assistants, like Ironman, with superpowers focused on improving and optimizing our health and everyday lives. The smart exosuit can provide an extra set of eyes that monitors the environment and warns of threats when we walk home at night. This exosuit, shaped as a knitted robot, could become an individual coach to help us perfect a tennis stroke or an assembly sequence. This is just a snapshot of a machine-enhanced future we can imagine. There are so many ways in which our lives can be augmented by robots and AI.

This positive human-machine relationship, in which machines are helpful assistants, is closer to my vision of the future than the scenarios in which the machines either take over as maniacal overlords or solve all of humanity's problems. This vision is starting to mature inside my lab, and in the labs of my friends and colleagues at other universities and institutions and some forward-thinking companies. This future does not resemble the dystopia depicted in so many books, movies, and articles. But none of us expects it to be a perfect world, either, which is why we design and develop the work with potential dangers in mind.

While AI is concerned with developing the science and engineering of intelligence for cognitive tasks, robotics is concerned with physical-world interactions by developing the science and engineering of autonomy. Specifically, robots are made of a body (hardware) and a brain (algorithms and software). For any task that requires machine assistance, we need bodies capable of doing the task and brains capable of controlling the bodies to do the task. The main tasks studied in robotics are mobility (navigating on the ground, in air, or underwater), manipulation (moving objects in the world), and interaction (engaging with other machines and with people).

We have already come a long way. Today's state of the art in robotics, AI, and machine learning is built on decades of advancements and has great potential for positive impact. The first industrial robot, called The Unimate, was introduced in

1961. It was invented to perform industrial pick and place operations. By 2020, the number of industrial robots in operation reached around twelve million, while the number of domestic robots reached thirty-one million.[1] These industrial robots are masterpieces of engineering capable of doing so much more than people can, yet these robots remain isolated from people on factory floors because they are large, heavy, and dangerous to be around. By comparison, organisms in nature are soft, safe, compliant, and much more dexterous and intelligent. Soft-bodied systems like the octopus can move with agility. The octopus can bend and twist continuously and compliantly to execute many other tasks that require dexterity and strength, such as opening the lid of a jar. Elephants can move their trunks delicately to pick up potato chips, bananas, and peanuts, and they can whip those same trunks with force enough to fight off a challenger. If robots could behave as flexibly, people and robots could work together safely side by side. But what would it take to develop robots with these abilities?

While the past sixty years have established the field of industrial robots and empowered hard-bodied robots to execute complex assembly tasks in constrained industrial settings, the next sixty years will usher in soft robots for human-centric environments and to help people with physical and cognitive tasks. While the robots of the past sixty years have mostly been inspired by the human form, shaped as industrial arms, humanoids, and boxes on wheels, the next phase for robots will include soft machines with shapes inspired by the animal kingdom and its diversity of forms, as well as by our own built environments. The new robot bodies will be built out of a variety of available materials: silicone, wood, paper, fabric, even food. These machines of our future have a broader range of customized applications.

Today's industrial manipulators enable rapid and precise assembly, but these robots are confined to operate independently from humans (often in cages) to ensure the safety of the humans around them. The lack of compliance in conventional actuation mechanisms is part of this problem. In contrast, nature is not fully rigid; it uses elasticity and compliance to adapt. Inspired by nature, soft robots have bodies made out of intrinsically soft and/or extensible materials (such as silicone rubbers or fabrics) and are safe for interaction with humans and animals. They have a continuously deformable structure with muscle-like actuation that emulates biological systems and provides them with a relatively large number of degrees of freedom as compared with their hard-bodied counterparts. Soft robots have capabilities beyond what is possible with today's rigid-bodied robots. For example, soft-bodied robots can move in more natural ways that include complex bending and twisting curvatures that are not restricted to the traditional rigid body kinematics of existing robotic manipulators. Their bodies can deform continuously, providing theoretically infinite degrees of freedom and allowing them to adapt their shape to their environments (such as by conforming to natural terrain or forming enveloping power grasps). However, soft robots have also been

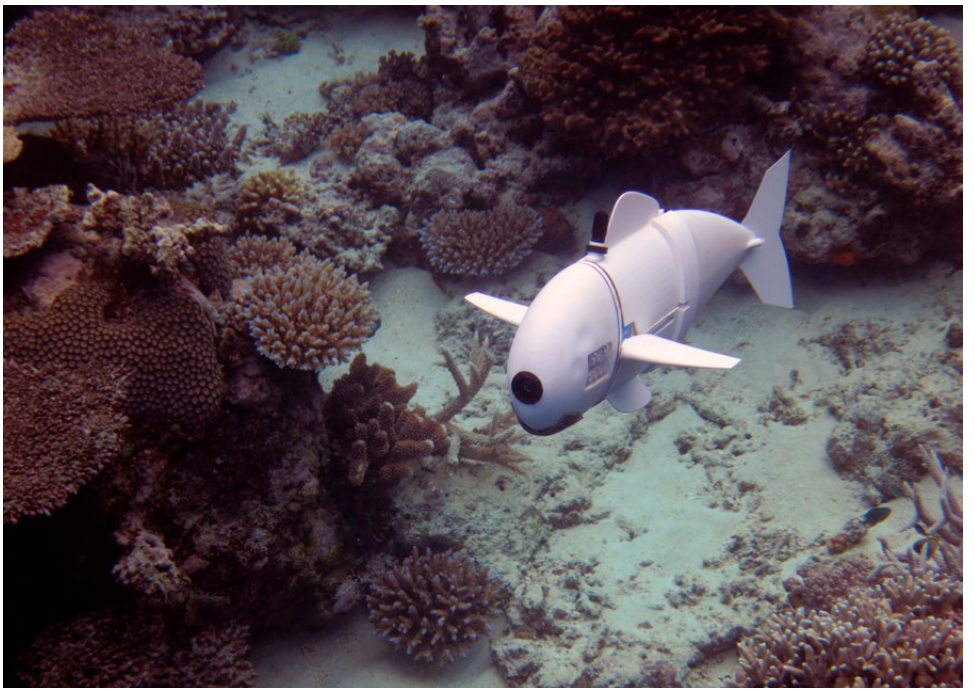
shown to be capable of rapid agile maneuvers and can change their stiffness to achieve a task- or environment-specific impedance.

What is soft, really? Softness refers to how stretchy and compliant the body of the robot is. Soft materials and electromechanical components are the key enablers for creating soft robot bodies. Young's modulus, which computes the ratio of stress to strain of a material when force is applied, is a useful measure of the rigidity of materials used in the fabrication of robotic systems. Materials traditionally used in robotics (like metals and hard plastics) have Young's moduli on the order of 10^9 to 10^{12} pascals (a unit of pressure), whereas natural organisms are often composed of materials (like skin and muscle tissue) with moduli on the order of 10^4 to 10^9 pascals, orders of magnitude lower than their engineered counterparts. We define soft robots as systems capable of autonomous behavior that are primarily composed of materials with moduli in the range of soft biological materials.

Current research on device-level and algorithmic aspects of soft robots has resulted in a range of novel soft devices. But how do we get to the point where soft robots deliver on their full potential? The capabilities of robots are defined by the tight coupling between their physical bodies and the computation that makes up their brains. For example, a robot fish must have both a body capable of swimming and algorithms to control its movement in water. Today's soft-bodied robots can do basic locomotion and grasping. When augmented with appropriate sensors and computation, they can recognize objects in restricted situations, map new environments, perform pick and place operations, and even act as a coordinated team.

Figure 1 shows SoFi, the soft robotic fish.[2] SoFi is an autonomous soft robot developed for close observations and interactions with marine life. SoFi enables people to observe and monitor marine life from a distance, without interference. The robot swims continuously at various depths in a biomimetic way by cyclic undulation of its posterior soft body. The fish controls the undulating motion of its tail using a hydraulically actuated soft actuator with two internal cavities separated by an inextensible constraint. The fish tail has two chambers with ribbed structure for pressurization, and the inextensible constraint is in the middle. Maneuvering is accomplished by moving water from one chamber to the other using a pump. When the pump moves water equally between the left and right chambers of the tail, the tail moves back and forth evenly, and the fish exhibits forward swimming. It is possible to make right-hand turns by pumping more water in the right chamber than the left and doing the reverse for left-hand turns. The swimming depth is controlled by two dive planes that represent the robot's fins. SoFi has onboard capabilities for autonomous operation in ocean environments, including the ability to move along 3D trajectories by adjusting its dive planes or by controlling its buoyancy. Onboard sensors perceive the surrounding environ-

Figure 1
SoFi, the Soft Robotic Fish for Underwater Observatories



SoFi, the soft robotic fish swimming in a coral reef. Source: Photo by Joseph DelPreto. See Robert K. Katzschmann, Joseph DelPreto, Robert MacCurdy, et al., "Exploration of Underwater Life with an Acoustically Controlled Soft Robotic Fish," *Science Robotics* 3 (16) (2018).

ment, and a mission control system enables a human diver to issue remote commands. SoFi achieves autonomy at a wide range of depths through 1) a powerful hydraulic soft actuator; 2) a control mechanism that allows the robot to adjust its buoyancy according to depth, thus enabling long-term autonomous operation; 3) onboard sensors to observe and record the environment; 4) extended ocean experiments; and 5) a mission control system that a human diver can use to provide navigation commands to the robot from a distance using acoustic signals. SoFi has the autonomy and onboard capabilities of a mobile underwater observatory, our own version of Jules Verne's marine observatory in *Twenty Thousand Leagues Under the Sea*. Marine biologists have long experienced the challenges of documenting ocean life, with many species of fish proving quite sensitive to the underwater movements of rovers and humans. While multiple types of robotic

instruments exist, the soft robots move by undulation and can more naturally integrate in the undersea ecosystems. Soft-bodied robots can move in more natural and quieter ways.

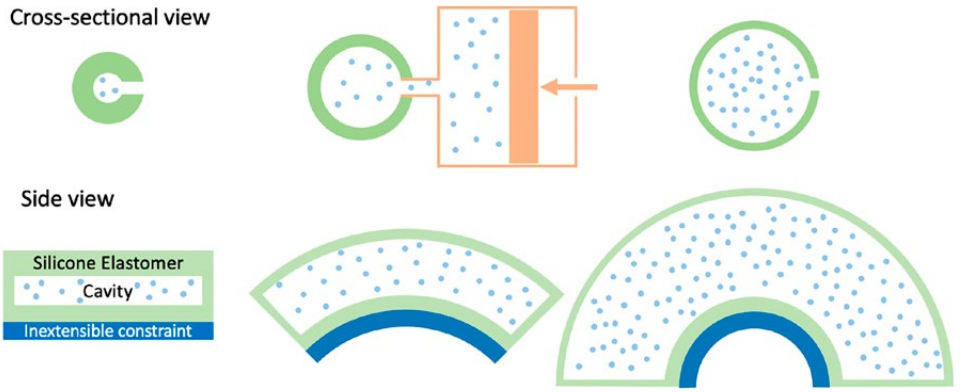
The body of a soft robot like SoFi may consist of multiple materials with different stiffness properties. A soft robot encases in a soft body all the subsystems of a conventional robot: an actuation system, a perception system, driving electronics, and a computation system, with corresponding power sources. Technological advances in soft materials and subsystems compatible with the soft body enable the autonomous function of the soft robot.

At the core of any robot is actuation. One of the primary focus areas to date for soft robots has been the exploration of new concepts for compliant yet effective actuators. Researchers have made progress on several classes of soft actuators, most prominently with fluidic or various electrically activated tendon actuators. Fluidic elastomer actuators (FEAs) are highly extensible and adaptable, low-power soft actuators. FEAs were used to actuate SoFi's tail. Figure 2 shows the actuation principle. A silicone chamber has an inextensible constraint. When it is pressurized – for example, with air or liquid – the skin expands and forms a curvature. By controlling this curvature, we can control the movement of the robot.

The soft actuator in Figure 2 can move along one axis and is thus called a one-degree-of-freedom actuator. Such an actuator can be composed in series and in parallel to create any desired compliant robotic morphology: a robotic elephant trunk, a robotic multifinger hand, a robotic worm, a robotic flower, a robotic chair, even a robotic lamp.

However, while achieving compliance, this FEA actuator structure has not achieved muscle-like or motor-like performance in terms of force, displacement, energy density, bandwidth, power density, and efficiency. In order to create muscle-like actuation, we can leverage the idea of combining soft bodies with compliant origami structures to act as "flexible bones" within the soft tissue. The idea of fluidic origami-inspired artificial muscles (FOAM) provides fluidic artificial muscles with unprecedented performance-to-cost ratio.[3] The FOAM artificial muscle system consists of three components: a compressible solid skeletal structure (an origami structure), a flexible fluid-tight skin, and a fluid medium. When a pressure difference is applied between the outside and the inner portion, a tension is developed in the skin that causes contraction that is mediated by the folded skeleton structure. In a FOAM system, the skin is sealed as a bag covering the internal components. The fluid medium fills the internal space between the skeleton and the skin. In the initial equilibrium state, the pressures of the internal fluid and the external fluid are equal. However, as the volume of the internal fluid changes, a new equilibrium is achieved. A pressure difference between the internal and external fluids induces tension in the flexible skin. This tension will act on the skeleton, driving a trans-

Figure 2
Soft Fluidic Actuation

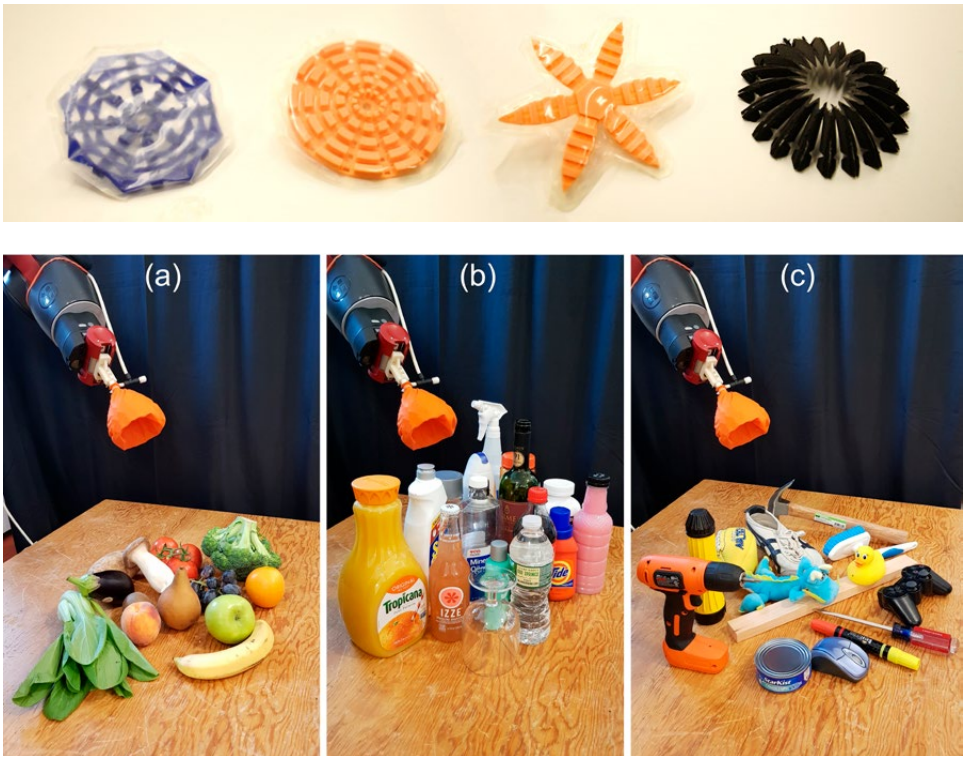


Source: Robert Katzschmann, "Building and Controlling Fluidically Actuated Soft Robots: From Open Loop to Model-based Control" (Ph.D. diss., MIT, 2013).

formation that is regulated by its internal skeletal geometry. These artificial muscles can be programmed to produce not only a single contraction, but also complex multiaxial actuation and even controllable motion with multiple degrees of freedom. Moreover, a variety of materials and fabrication processes can be used to build the artificial muscles with other functions beyond basic actuation. Experiments reveal that these muscles can contract over 90 percent of their initial lengths, generate stresses of approximately 600 kilopascals, and produce peak power densities over 2 kilowatts per kilogram: all equal to, or in excess of, natural muscle. For example, a 3 gram FOAM actuator that includes a zig-zag pattern for its bone structure can lift up to 3 kilograms! This architecture for artificial muscles opens the door to rapid design and low-cost fabrication of actuation systems for numerous applications at multiple scales, ranging from miniature medical devices to wearable robotic exoskeletons to large deployable structures for space exploration.

The soft FOAM grippers shown in Figure 3 are made from a soft origami structure, encased by a soft balloon.[4] When a vacuum is applied to the balloon, the origami structure – a design based on a folding pattern – closes around the object, and the gripper deforms to the geometric structure of the object. While this motion lets the gripper grasp a much wider range of objects than ever before, such as soup cans, hammers, wine glasses, drones, even a single broccoli floret or grape, the greater intricacies of delicacy – in other words, how hard to squeeze – require adding sensors to the gripper. Tactile sensors can be made from latex "bladders"

Figure 3
FOAM Grippers *(top)* and Objects that can be Handled with the Magic Origami Ball (Tulip) Gripper *(bottom)*



Source : Photos by Shuguang Li.

(balloons) connected to pressure transducers. The new sensors let the gripper not only pick up objects as delicate as potato chips, but it also classifies them, providing the robot with a better understanding of *what* it is picking up, while also exhibiting that light touch. When the embedded sensors experience force or strain, the internal pressure changes, and this feedback can be used to achieve a stable grasp.

In addition to such discrete bladder sensors, we can also give the soft robot bodies sensorized "skin" to enable them to see the world by feeling the world. The sensorized skin provides feedback along the entire contact surface, which is valuable for learning the type of object it is grasping and exploring the space of the robot through touch. Somatosensitive sensors can be embedded in the silicone body of the robot using 3D printing with fugitive and embedded ink. Alternatively, elec-

trically conductive silicone can be cut using a variety of stretchable kirigami patterns and used for the sensor skin of the robot. Machine learning can then be used to associate skin sensor values with robotic deformations, leading to proprioceptive soft robots that can "see" the world through touch.

The robot body needs a robot brain to command and coordinate its actions. The robot brain consists of the set of algorithms that can get the robot to deliver on its capabilities. These algorithms typically map onto computation for physically moving the components of the robot (also called low-level control) and computation for getting the robot to perform its assignment (also called high-level or task-level control).

While we have a surge in developing soft bodies for robots, the computational intelligence and control of these robots is more challenging. Results from rigid robots do not immediately translate to soft robots because of their inherent high dimensionality. The state of a rigid robot can be described compactly with a finite set of degrees of freedom: namely, the displacement of each of its joints as described in their local coordinate frames. Their bodies are constrained by the inflexible nature of their rigid links. Fully soft robots, by contrast, may not have a traditional joint structure, relying on their flexible body to solve tasks. Soft robots have a dramatically different interaction with the environment through rich compliant contact. There is currently a divide in the approach to control: rigid robots control contact forces/contact geometry while soft robots rely almost entirely on open-loop interactions, mediated by material properties, to govern the resulting forces/geometry. One strategy for bridging this gap lies in optimization-based control via approximate dynamic models of the soft interface: models with a fidelity that is customized to the task. The governing equations of the soft robots are complex continuum mechanics formulations that are typically approximated using high-dimensional finite-element methods. The dynamics are highly nonlinear, and contacts with the environment make them nonsmooth. These models are too complex for state-of-the-art feedback design approaches, which either make linearity assumptions or scale badly with the size of the state space. The challenge is to find models simple enough to be used for control, but complex enough to capture the behavior of the system.

For low-level control of soft robots, we can often identify a sequence of actuated segments, in which torques are dominant, so it is possible to assume the curvature to be constant within each segment, leading to a finite-dimensional Piecewise Constant Curvature (PCC) kinematic description. We can then describe the PCC of the soft robot through an equivalent rigid robot with an augmented state space.

Task-level control of soft robots is often achieved in a data-driven way using machine learning. Some of today's greatest successes of machine learning are due to a technique called *deep learning*. Deep learning uses data – usually millions of

hand-labeled examples – to determine the weights that correspond to each node in a convolutional neural network (CNN), a class of artificial neural networks, so that when the network is used with new input, it will classify that input correctly. Deep learning has been successfully applied to soft robots to provide them with capabilities for proprioception (sensitivity to self-movement, position, and action), exteroception (sensitivity to outside stimuli), and grasping.

But deep learning faces a number of challenges. First among them is the data. These techniques require data availability, meaning massive data sets that have to be manually labeled and are not easily obtained for every task. The quality of that data needs to be very high, and it needs to include critical corner cases – that is, cases outside the training distribution or outside usual operations – for the application at hand. If the data are biased, the performance of the algorithm will be equally bad. Furthermore, these systems are black boxes: there is no way for users of the systems to truly "learn" anything based on the system's workings. It is difficult to detect behavior that is abnormal from a safety point of view. As a result, it is hard to anticipate failure modes tied to rare inputs that could lead to potentially catastrophic consequences. We also have robustness challenges and need to understand that the majority of today's deep-learning systems perform pattern matching rather than deep reasoning. Additionally, there are sustainability issues related to data-driven methods. Training and using models consume enormous amounts of energy. Researchers at the University of Massachusetts Amherst estimated that training a large deep-learning model produces 626,000 pounds of carbon dioxide, equal to the lifetime emissions of five cars. The more pervasive machine learning becomes, the more of these models will be needed, which in turn has a significant environmental impact.

Today's machine learning systems are so costly because each one contains hundreds of thousands of neurons and billions of interconnections. We need new ideas to develop simpler models, which could drastically reduce the carbon footprint of AI while gaining new insights into intelligence. The size of a deep neural network constrains its capabilities and, as a result, these networks tend to be huge and there is an enormous cost to running them. They are also not interpretable. In deep neural networks, the architecture is standardized, with identical neurons that each compute a simple thresholding function. A deep neural network that learns end-to-end from human data how to control a robot to steer requires more than one hundred thousand nodes and half a million parameters.

Using inspiration from neuroscience, my colleagues and I have developed neural circuit policies,[5] or NCPs, a new approach to machine learning. With NCPs, the end-to-end steering task requiring more than one hundred thousand simple neurons can be learned with nineteen NCP neurons in the deep neural network model, resulting in a more efficient and interpretable system. The neuroscience inspiration from the natural world is threefold. First, NCP neurons can compute more than a

step function; each NCP neuron is a liquid time differential equation. Second, NCP neurons can be specialized, such as input, command, and motor neurons. Third, the wiring architecture has organism-specific structure. Many other tasks related to spatial navigation and beyond can be realized with neuroscience-inspired, compact, and interpretable neural circuit policies. Exploring robot intelligence using inspiration from the natural world will yield new insights into life and provide new computational models for intelligence that are especially useful for soft robots.

Novel soft design, fabrication, and computation technologies are ushering in a new era of robots that come in a variety of forms and materials and are designed to help people with physical tasks in human-centric environments. These robots are smaller, safer, easier to fabricate, less expensive to produce, and more intuitive to control.

Robots are complex systems that tightly couple the physical mechanisms (the body) with the software aspects (the brain). Recent advances in disk storage, the scale and performance of the Internet, wireless communication, tools supporting design and manufacturing, and the power and efficiency of electronics, coupled with the worldwide growth of data storage, have helped shape the development of robots. Hardware costs are going down, the electromechanical components are more reliable, the tools for making robots are richer, the programming environments are more readily available, and the robots have access to the world's knowledge through the cloud. Sensors like the LiDAR (light detection and ranging) systems are empowering robots to measure distances very precisely. Tiny cameras are providing a rich information stream. Advances in the development of algorithms for mapping, localization, object recognition, planning, and learning are enabling new robotic capabilities. We can begin to imagine the leap from the personal computer to the personal robot, leading to many applications in which robots exist pervasively and work side by side with humans.

How might these advances in robotics shape our future? Today, if you can think it, you can write it on paper. Imagine a word where if you can think it, you can make it. In this way, the scientific advancement of soft robotics could give every one of us superpowers. Each of us could use our talents, our creativity, and our problem-solving skills to dream up robots that save lives, improve lives, carry out difficult tasks, take us places we cannot physically go, entertain us, communicate, and much more. In a future of democratized access to robots, the possibilities for building a better world are limitless. Broad adoption of robots will require a natural integration of robots in the human world, rather than an integration of humans into the machines' world.

These machines from our future will help us transform into a safer society living on a healthier planet, but we have significant technological and societal challenges to get to that point.

On the technical side, it is important to know that most of today's greatest advances in machine learning are due to decades-old ideas enhanced by vast amounts of data and computation. Without new technical ideas and funding to back them, more and more people will be ploughing the same field, and the results will only be incremental. We need major breakthroughs if we are going to manage the major technical challenges facing the field. We also need the computational infrastructure to enable the progress, an infrastructure that will deliver to us data and computation like we get water and energy today: anywhere, anytime, with a simple turn of a knob. And we need the funding to do this.

On the societal side, the spread of AI and robots will make our lives easier, but many of the roles that they can play will displace work done by humans today. We need to anticipate and respond to the forms of economic inequality this could create. In addition, the lack of interpretability and dependence could lead to significant issues around trust and privacy. We need to address these issues, and we need to develop an ethics and legal framework for how to use AI and robots for the greater good. As we gather more data to feed into these AI systems, the risks to privacy will grow, as will the opportunities for authoritarian governments to leverage these tools to curtail freedom and democracy in countries around the world.

These problems are not like the COVID-19 pandemic: we know they are coming, and we can set out to find solutions at the intersection of policy, technology, and business, in advance, now. But where do we begin?

In its report on AI ethics, the Defense Innovation Board describes five AI principles. First is responsibility, meaning that humans should exercise appropriate levels of judgment and remain responsible for the development, deployment, use, and outcomes of these systems. Second, equitability, meaning that we need to take deliberate steps to anticipate and avoid unintended bias and unintended consequences. Third is traceability, meaning that the AI engineering discipline should be sufficiently advanced such that technical experts possess an appropriate understanding of the technology, development processes, and operational methods of its AI systems. Fourth is reliability, meaning that AI systems should have an explicit, well-defined domain of use, and the safety, security, and robustness of such systems should be tested and assured. And finally, governance, meaning that AI systems should be designed and engineered to fulfill their intended function, while possessing the ability to detect and avoid unintended harm or disruption.[6] Beyond these general principles, we also need to consider the environmental impacts of new technologies, as well as what policy actions are needed to stem possible dangers associated with technological advances.

Neural circuit policies may sound like phrases you would only ever hear walking the hallways of places like CSAIL, the Computer Science and Artificial Intelligence Laboratory at MIT, where I work. We do not need everybody to understand in great detail how this technology works. But we do need our policy-makers and

citizens to know about the effects of new technologies so we can make informed decisions about their adoption. Together, we can build a common understanding around five vital questions: First, what *can* we do, or more specifically, what is really possible with technology? Second, what *can't* we do, or what is not yet possible? Third, what *should* we do? Fourth, what *shouldn't* we do? There are technologies and applications that we should rule out. And finally, what *must* we do. I believe we have an obligation to consider how AI technology can help. Whether you are a technologist, a scientist, a national security leader, a business leader, a policy-maker, or simply a human being, we all have a moral obligation to use AI technology to make our world, and the lives of its residents, safer and better, in a just, equitable way.

The optimist in me believes that can and will happen.

AUTHOR'S NOTE

The author gratefully acknowledges the following support: NSF grant No. EFRI-1830901, the Boeing Company, JMPC, and the MIT-Air Force AI Accelerator program.

ABOUT THE AUTHOR

Daniela Rus, a Fellow of the American Academy since 2017, is the Andrew (1956) and Erna Viterbi Professor of Electrical Engineering and Computer Science and the Director of the Computer Science and Artificial Intelligence Laboratory (CSAIL) at MIT, and the Deputy Dean of Research in MIT's Schwarzman College of Computing. She is also a Fellow of the National Academy of Engineering. She has published in such journals as *Nature*, *Proceedings of the National Academy of Sciences*, and *Science*, and in professional journals such as *International Journal of Robotics Research* and *IEEE Transactions on Robotics*.

ENDNOTES

[1] International Federation of Robots, https://ifr.org.

[2] Daniela Rus and Michael T. Tolley, "Design, Fabrication and Control of Soft Robots," *Nature* 521 (7553) (2015): 467–475.

[3] Shuguang Li, Daniel M. Vogt, Daniela Rus, and Robert J. Wood, "Fluid-Driven Origami Inspired Artificial Muscles," *Proceedings of the National Academy of Sciences* 114 (50) (2017): 13132–13137.

[4] Shuguang Li, John J. Stampfli, Helen J. Xu, et al., "A Vacuum-Driven Origami 'Magic-Ball' Soft Gripper," *2019 International Conference on Robotics and Automation (ICRA)* (2019): 7401–7408, https://doi.org/10.1109/ICRA.2019.8794068.

[5] Mathias Lechner, Ramin Hasani, Alexander Amini, et al., "Neural Circuit Policies Enable Auditable Autonomy," *Nature Machine Intelligence* 2 (10) (2020): 642–652.

[6] Defense Innovation Board, *AI Principles: Recommendations on the Ethical Use of Artificial Intelligence by the Department of Defense* (Washington, D.C.: U.S. Department of Defense, 2019).

Multi-Agent Systems: Technical & Ethical Challenges of Functioning in a Mixed Group

Kobi Gal & Barbara J. Grosz

In today's highly interconnected, open-networked computing world, artificial intelligence computer agents increasingly interact in groups with each other and with people both virtually and in the physical world. AI's current core challenges concern determining ways to build AI systems that function effectively and safely for people and the societies in which they live. To incorporate reasoning about people, research in multi-agent systems has engendered paradigmatic shifts in computer-agent design, models, and methods, as well as the development of new representations of information about agents and their environments. These changes have raised technical as well as ethical and societal challenges. This essay describes technical advances in computer-agent representations, decision-making, reasoning, and learning methods and highlights some paramount ethical challenges.

For many decades after its inception, AI's most pressing question, its core challenge, was to determine whether it was possible to build computer systems able to perform intelligent behaviors like engaging in a conversation, playing chess, or fixing a complex piece of machinery. By the twenty-first century, the use of computer systems had evolved from a single person with computing expertise interacting with a single system to a highly interconnected, open-networked computing world in which people's online activities connect them instantly with many different systems and people. There are thus ever more situations in which AI agents interact in groups with each other and with people both virtually and in the physical world. AI's most pressing questions today – its core challenges – center on determining ways to build AI systems that function effectively and safely for people and the societies in which they live. Concomitantly, research in the multi-agent systems area of AI increasingly addresses challenges of building capabilities for AI agents to act effectively in groups that include people: for instance, investigating robot-human collaborations in industrial settings, coordinating health care for patients seeing multiple providers, and adapting educational content to individual students' needs. We refer to these as *mixed-agent groups*.

© 2022 by Kobi Gal & Barbara J. Grosz
Published under a Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) license
https://doi.org/10.1162/DAED_a_01904

AI research traditionally modeled the behavior of an individual computer agent, whether embodied in a physical system (such as robots) or embedded in a software system (such as recommendation systems or customer service chatbots), as an act-observe-update-decide cycle: the agent does something in its world, observes the ways that world changes, revises its beliefs about the world based on those observations, and determines what action, if any, to take next. Some AI agent models determine next actions based on maximizing a utility function, while others reason logically. These individual-agent models have regarded other agents, whether computer agents or people, as part of the agent's environment. To enable agents to participate effectively in mixed-agent groups required two significant modeling changes: the design of ways to represent the mental state of other agents and the development of models of human decision-making and communication capacities that respect the complementarities of human and computer-agent capabilities. For instance, computer systems have vastly greater ability than humans to access and summarize large amounts of data, while people's capabilities for causal and counterfactual reasoning far outstrip those of AI systems.

Mental state representations enable computer agents to treat other agents (whether human or computer) as full-fledged actors that have beliefs and abilities to make decisions, to act on those decisions, and to reason about the beliefs and actions of other agents in their environment. Computer agents can then recognize ways that actions of one agent may affect the beliefs and influence subsequent actions of other agents. Research on standard multi-agent models, including both logic-based belief-desire-intention models and probabilistic Markov decision process models, has generated a variety of techniques for multi–computer agent groups, for both competitive and cooperative settings, yielding a diverse range of successfully deployed systems.[1]

To develop realistic models of human decision-making has required changes to every component of the traditional act-observe-update-decide cycle. AI researchers have developed new models, methods, and agent designs that incorporate reasoning about people for both machine-learning–based systems and logic-based systems. While agents in mixed-agent groups, like those in multi-agent systems generally, might compete, the focus of research has been on settings in which computer agents cooperate or fully collaborate with people in their mixed-agent group. These changes have raised not just new technical challenges, but also paramount ethical and societal-impact challenges.

Research on AI models of collaboration laid the foundations for reasoning about people as participants in mixed-agent groups.[2] These models stipulate as a defining characteristic of collaboration that all team participants share an overarching goal. The models provide theoretical frameworks for representing and reasoning about the mental state and communication requirements

for successful teamwork activities. Related work in AI and cognitive science specifies the obligations collective intentionality entails.[3]

Significant recent research focuses on settings in which computer agents need to coordinate with people, but absent an overall teamwork goal. For instance, an autonomous vehicle does not share an overarching destination goal with other drivers or pedestrians it encounters; the autonomous vehicle and others on the road do not make a team. Aspects of the early frameworks are also relevant to such settings, as is early work specifying the roles of social norms in coordinating behavior of multiple agents without a shared goal who nonetheless need to avoid conflict.[4] Key insights of this early work include establishing the need to explicitly design agents for collaboration, showing that the requisite capabilities could not be patched on, and the need for revisions of plan representations and decision-making algorithms for them.[5]

Subsequent work in both logical and machine learning paradigms has demonstrated the benefits of developing algorithms that consider the combined performances of people and agents rather than focusing on the autonomous performance of a computer agent in isolation.[6] For example, methods that optimize for agents to complement human capabilities or to balance human and computer agent preferences outperformed individual human and computer performances.[7] Other work deploys cross-training to improve human-robot team performance.[8] A consensus is emerging from this research of the importance of bringing insights from the social sciences to bear in designing agents for working with people.[9]

The advent of large-scale Internet activities – from citizen science to online learning and question-and-answer sites – has provided researchers with significantly more data than ever before about people's behaviors and preferences, creating new technical opportunities and raising new AI research questions. Not only do people's decision-making processes often not adhere to standard assumptions about optimizing for utility, but these larger-scale settings require computer agents to operate in the "open world," rather than in well-defined, constrained, and therefore more easily specifiable environments ("closed worlds").[10] As a result, agent designs need to accommodate both *scale* – a significant increase in the number of people an agent may work with – and operating "in the wild": that is, in open worlds in which computer agents have only partial information about other agents and much less control. Further challenges arise from the need for computer-agent behaviors and explanations to mesh with people's expectations.[11]

We briefly describe AI researchers' advances on three core computer-agent capabilities that are enabling agents to participate more effectively in mixed-agent groups: 1) *decision-making* about what to do next, considering the potential effects of an agent's actions on other agents' beliefs and decision-making, as well as on the environment; 2) *reasoning* to draw conclusions about the effects of an agent's actions on that environment, including any causal connections; and 3) *learning* from

the effects it observes in the environment and on others' actions. This research has led to paradigmatic shifts in a variety of AI methods and algorithms, as well as to the development of *new representations* of information about agents and the environments in which they act.

New representations of actions, plans, and agent interactions enable agents to reason about their human partners despite having limited information about their beliefs and capabilities. For instance, a digital personal assistant may not know which route a person is taking to get home, and a health care coordination system may need to learn interaction patterns among medical providers as they evolve.

Novel ways of representing task and plan knowledge – for instance, with Ece Kamar, we expanded the SharedPlans specification of teamwork – enable collaboration when an agent does not know which of several plans a person is following.[12] To enable computer agents to reason effectively about information sharing when they lack *a priori* knowledge of other agents' plans (as required by standard information-sharing algorithms), Ofra Amir and colleagues developed a representation of "mutual influence potential" networks for teams that operate over long periods of time (such as project management and health care teams).[13] To address the need for computer-agent collaborators to adapt to their human partners' actions, Stefanos Nikolaidis and colleagues developed a representation for Markov decision processes that evolves through cross-training, and they demonstrated that cross-training outperforms other training regimes.[14]

New methods of decision-making have been designed by AI researchers to reason about social influences on people's behavior in negotiation; to determine when to share information with partners in a group activity; and, for large-scale groups, to identify the best people for certain tasks and to provide incentives for them to contribute to group activities.

For computer agents to negotiate effectively with people, they need to take into account findings in the social sciences that have revealed social influences on people's negotiation strategies. Research incorporating such findings into agent negotiation strategies – by representing social attributes in the decision-making model – has demonstrated the ability of such socially aware agents to reach agreements that benefit all participants. For instance, through empirical investigations, we showed that people's willingness to accept offers is affected by such traits as altruism and selfishness, and that agents incorporating these traits into their negotiation strategies outperform traditional game-theoretic equilibria strategies.[15] Amos Azaria and colleagues improved agent success in advising a person on the best route to a destination by incorporating a model of people's behavior in repeated negotiations.[16] And Arlette van Wissen and colleagues found that although

people trust computer agents as much as other people in negotiations, they treat them less fairly.[17] Agents negotiating for people also need to model their preferences. For example, an agent might assist a consumer in negotiating the best deal for an item available from multiple online sellers who offer similar products at varying prices and characteristics – used or new, full or delayed payment – saving the consumer time and money. If the consumer is price sensitive, the agent could negotiate a lower price while agreeing to make the payment in advance.

To coordinate their activities, participants in mixed-agent groups typically must share information with each other about their activities, environments, and tasks. Decisions about what information to share, and when, are more complicated when computer agents are not privy to important task-related information that people hold. For example, a driver-assist system considering whether to alert the driver to unexpected traffic ahead on a possible route that allowed for a side-trip to a pharmacy may not be certain about the driver's current preferences with respect to making that stop. As a result, it may not know if this traffic situation is on the route the driver is taking and thus whether notifying the driver would be useful or an unnecessary interruption. Information exchanges – whether an unneeded alert or a request for irrelevant information – generate cognitive and communication costs. Research on managing information exchange to avoid overburdening people includes theoretical model development and empirical studies.

With Ece Kamar, we identified the class of "nearly decomposable" settings, in which computer agents need to reason about only that subset of their human partners' actions that interact with the agent's actions.[18] We developed a multi-agent Markov decision process for such settings that enables more efficient inference for interruption management. An empirical study using this method identified factors influencing people's acceptance of an agent's interruptions.

In work on information sharing for team settings in which agents have very limited information about their human partners, Ofra Amir and colleagues developed an algorithm that identifies the information that is most relevant to each team member using the influence potential networks described earlier.[19] The results of a laboratory study using this algorithm demonstrated that information-sharing decisions based on the influence-potential representation yielded higher productivity and lower perceived workload compared with standard human-computer interaction approaches.

In such large-scale settings as disaster response and online forums, the standard multi-agent systems' role assignment problem – the problem of identifying the best agent for a particular task – is more difficult because less information is directly available about (human) participants' capabilities. These settings also introduce a new role-assignment challenge : namely, keeping people engaged.

Methods that integrate behavior prediction into decision-making processes enable inferring people's capabilities from their prior interactions and thus pre-

dicting the best person to assign a task to. Research on engagement includes the use of reinforcement learning to generate motivational messages.[20] The benefits of these approaches have been demonstrated in citizen science applications such as classifying celestial bodies and identifying EEG patterns.[21]

Reasoning and learning are tightly coupled. We discuss them together because new methods developed to jointly learn and use models of people's behavior have been consequential for mixed-agent group settings. Important new reasoning capabilities include 1) methods for predicting people's behavior from data about the actions they have taken in the past, their causal effects, and the outcomes that result; 2) techniques for agents to use advice and feedback from people to learn more effectively; and 3) methods for agents to explain their choices and recommendations well enough that people understand them. For example, to find the sequence of math problems that maximizes students' learning gains, an AI tutor needs to predict their responses to math problems. It also needs to be able to explain its problem choices to students, and possibly their teachers.[22]

Computer agents in mixed-agent groups need to model people's past actions and to predict their likely future actions. Machine learning algorithms face a compatibility-performance trade-off: updating machine learning systems with new data may improve their overall performance, but the updated predictions may decrease trust in the system by individuals for whom the predictions no longer work. To address this problem, Jonathan Martinez and colleagues defined machine learning algorithms that personalize their updates to individual users, which not only yields higher accuracy but also makes models more compatible with people's expectations.[23] They established the efficacy of this approach empirically by comparing it with a baseline method that did not personalize the model's updates.

People "in the wild" also make computer agents' plan recognition – the ability to determine what others are doing and why – more difficult, since they often exhibit complex planning behaviors: they may follow multiple plans, interleave actions from different plans, or perform actions that are redundant, wrong, or arbitrary. Novel plan and goal recognition algorithms have been developed to enable agents to adapt to people's exploratory and error-prone behavior. They use various techniques, including heuristics and approaches that replace predefined libraries of possible plans with generating plans on the fly.[24] To enable agents to support people's understanding of plans of other agents (human and computer) in their groups, researchers have designed new types of visualizations for presenting inferred plans to people in ways that facilitate their understanding of others' plans.[25]

Reinforcement learning algorithms enable agents to learn about their environment and about other agents through exploration and trial and error. Mixed-agent groups introduce a new possibility: algorithms can incorporate guidance and feedback from people who have relevant task expertise or knowledge of the

agent's environment and thus significantly facilitate agent learning. W. Bradley Knox and Peter Stone combined feedback from human teachers, who give positive or negative signals to the agent trainee, with autonomous learning about the environment.[26] Travis Mandel and colleagues augmented a reinforcement algorithm with a method for querying people about the best action to perform.[27] Their empirical studies demonstrated significant improvements to algorithm performance for domains with large numbers of actions. Matthew E. Taylor and colleagues showed that agents could adapt a policy to a new domain more effectively if a person first demonstrates how to act in that domain.[28] In this work, short episodes of human demonstrations led to rapid savings in learning time and policy performance for agents in different robot soccer simulation tasks.

For people to trust agents, the models they use to predict people's behavior not only need to perform well according to machine learning systems' metrics, but also to produce interpretable predictions – their action choices need to make sense to the people with whom they interact.[29] As all applications of AI machine learning methods have this need for "interpretability," a variety of research studies have investigated the design of "interpretable models" as well as ways to measure the interpretability of machine learning models in practice.[30]

The evaluation of multi-agent systems becomes significantly more complicated when an agent group includes people. Testing in the wild – that is, in the actual intended situations of use – may be costly both practically and ethically. In response to this challenge, researchers have developed various testbed systems that enable initial evaluation of effectiveness of computer-agent decision-making algorithms in lab (or lab-like) settings. They enable testing of new methods on intended user populations without such costs, allowing agent designers to better determine responses to agents' decisions as well as to compare the performance of different computational decision-making strategies. Some testbed systems have also been used to gather information about people's decision-making strategies to help improve the performance of learning algorithms.

Colored Trails, one of the first such testbeds, enabled the development of a family of games that facilitated the analysis of decision-making strategies, including negotiation strategies and coalition formation in widely varying settings.[31] The Genius testbed (General Environment for Negotiation with Intelligent multi-purpose Usage Simulation) advances research on bilateral multi-issue negotiation by providing tools for specific negotiation scenarios and negotiator preference profiles and for computing and visualizing optimal solutions.[32] The IAGO testbed (Interactive Arbitration Guide Online) provides a web-based interaction system for two-agent bargaining tasks. It has been used to study the role of affect and deception on negotiation strategies in mixed-agent groups.[33] Both Genius and IAGO testbeds have been used in competitions that compare computational strategies for negotiating with people.[34]

Research and development of computer agents capable of participating effectively in mixed-agent groups raise various ethical issues. Some are inherited from AI generally: for instance, avoiding bias, ensuring privacy, and treating people's data ethically. Others result from the mixed-agent group setting entailing that people and computer agents work together and, in some cases, share decision-making. Further, computer agents may be designed to influence people's behavior, make decisions related to people's futures, and negotiate successfully with people. While the roles computer agents and people assume vary within and across application domains, that people are inherent to the definition of "mixed-agent group" makes addressing particular ethical challenges of the utmost importance. We briefly discuss three challenges mixed-agent group research raises, all of which will require research done in concert with social and behavioral scientists and ethicists. We note that choices among ethical values and setting of norms are responsibilities of the societies in which these agent systems are used. Our discussion of ethical challenges thus presumes norms are established by communities of use, policy-making organizations, governmental bodies, or similar entities external to the research effort.

Challenge 1: Inclusive design and testing. The testing of new mixed-agent group algorithms and systems must involve the full range of people expected to participate in group undertakings with such agents. Further, whether for research or for system development, in designing mixed-agent group agents to align with societal values, designers must consider and engage at all stages of the work with the full spectrum of people with whom these agents are intended to interact. For instance, in the initial design stage, researchers should conduct informative interviews or observations to determine system goals and characteristics appropriate for the intended user population.[35]

Inclusivity generates particular challenges when designing new representations, whether models are explicitly designed or derived by machine learning methods. For instance, when developing new representations of tasks and plans, designers need to engage not only the kinds of people agents are likely to work with on a task, but also the kinds of people potentially affected by agent actions and decisions: for example, in a health care setting, the design of an agent that will work with physicians, nurses, and patients, as well as hospital administrative staff, should include physicians, nurses, and patients in the design cycle.

The need for inclusivity at the design stage also arises in areas of learning and reasoning. For example, when developing models of people's behavior, it is crucial for agents to handle adequately all types of people whose behavior it may need to track.

Challenge 2: Avoiding deception and exploitation. The use of social science factors in negotiation algorithms or for behavior modification (like nudges) may have purposes that engender unethical behavior. Mixed-agent group work on negoti-

ation may raise significant questions if the negotiation algorithm focuses only on improving the computer agent's outcome and deploys deception, rather than balancing good for all parties.[36] Similarly, role assignment in some ride-sharing applications has raised significant questions of deception and exploitation.

For agents in mixed-agent groups to be trustworthy, any use of deceptive strategies must be revealed. Researchers developing and deploying negotiation and behavior modification strategies must explain the rationale for them and make evident the ethical challenges they raise for any system that deploys them in applications and possible mitigations.

Challenge 3: Preventing or mitigating unanticipated uses of models and algorithms. The development of new representations and algorithms (such as for information sharing, role assignment, or behavior modeling) is typically driven by an intended application. The resulting learned representations and models may not be appropriate for other applications or may have consequences that were not anticipated when design was focused on the initial intended application. For example, a ride-sharing company might decide to adopt one of the "motivational" algorithms developed in the context of citizen science to attempt to keep drivers working when the system predicts they are close to quitting for the day. While there may be no serious downsides to encouraging someone to continue working on a science project despite being tired, there can be serious consequences from drivers working when fatigued. In some cases, the technology may be sufficiently unreliable or human oversight may be sufficiently inadequate that the unanticipated use should not be allowed. Researchers, system designers, and developers all bear responsibility for preventing the misuse of these technologies.

As mixed-agent groups become the norm in ever more multi-agent domains, advances in multi-agent systems research provide foundations for developing computer agents able to be effective partners in such settings. This work has also revealed a variety of new research challenges and raised important questions of ethical and societal impact.

For these reasons and others, successes in laboratory settings have not yet been translated into deployed systems on a large scale. The inadequacies of automated call centers and the difficulties Amazon fulfillment center workers have experienced working with robots illustrate the problems that arise when computer agents' activities do not mesh well with their human coworkers'. Perhaps the greatest challenge of developing computer agents technically and ethically adequate for participation in mixed-agent group undertakings is to fully recognize the sociotechnical nature of such activities. This recognition should lead not only to different kinds of algorithms, but also to processes for system development and deployment that take account of human capabilities, societal factors, and human-computer interaction design principles.

These challenges do not belong to research alone. If AI systems are to function effectively and safely for people and the societies in which they live, they require attention through the full pipeline from design through development, testing, and deployment. Addressing these challenges is all the more important given the recent broad range of national-level calls for developing effective methods for human-centered AI and for human-AI collaborations.

ABOUT THE AUTHORS

Kobi Gal is an Associate Professor in the Department of Software and Information Systems Engineering at the Ben-Gurion University of the Negev, and a Reader in the School of Informatics at the University of Edinburgh. His contributions to artificial intelligence include novel representations and algorithms for autonomous decision-making in heterogeneous groups comprising people and computational agents. They have been published in various highly refereed venues in artificial intelligence and in the learning and cognitive sciences. Among other awards, Gal is the recipient of the Wolf Foundation's Krill Prize for Israeli scientists and a Marie Curie European Union International Fellowship. He is a Senior Member of the Association for the Advancement of Artificial Intelligence.

Barbara J. Grosz, a Fellow of the American Academy since 2004, is the Higgins Research Professor of Natural Sciences at the John A. Paulson School of Engineering and Applied Sciences at Harvard University. Her contributions to the field of artificial intelligence include fundamental advances in natural language dialogue processing and in theories of multi-agent collaboration and their application to human-computer interaction, as well as innovative uses of models developed in this research to improve health care coordination and science education. She cofounded Harvard's Embedded Ethics program, which integrates teaching of ethical reasoning into core computer science courses. Grosz is also known for her role in the establishment and leadership of interdisciplinary institutions and for her contributions to the advancement of women in science. She is a member of the American Philosophical Society and the National Academy of Engineering, an elected fellow of several scientific societies, and recipient of the 2009 ACM/AAAI Allen Newell Award, the 2015 IJCAI Award for Research Excellence, and the 2017 Association for Computational Linguistics Lifetime Achievement Award.

ENDNOTES

[1] David C. Parkes and Michael P. Wellman, "Economic Reasoning and Artificial Intelligence," *Science* 349 (6245) (2015): 267–272; and Michael Wooldridge, *An Introduction to Multiagent Systems* (Hoboken, N.J.: John Wiley & Sons, 2009).

[2] Barbara J. Grosz and Sarit Kraus, "Collaborative Plans for Complex Group Action," *Artificial Intelligence* 86 (2) (1996): 269–357; Hector J. Levesque, Philip R. Cohen, and José H. T. Nunes, "On Acting Together," in *Proceedings of the Eighth National Conference on Artificial Intelligence* (Menlo Park, Calif.: AAAI Press, 1990); and David Kinny, Elizabeth Sonenberg, Magnus Ljungberg, et al., "Planned Team Activity," in *Artificial Social Systems*, ed. Cristiano Castelfranchi and Eric Werner (New York: Springer, 1992).

[3] Rosaria Conte and Cristiano Castelfranchi, *Cognitive and Social Action* (London: UCL Press, 1995); and Barbara J. Grosz and Luke Hunsberger, "The Dynamics of Intention in Collaborative Activity," *Cognitive Systems Research* 7 (2–3) (2006): 259–272.

[4] Yoav Shoham and Moshe Tennenholtz, "On the Emergence of Social Conventions: Modeling, Analysis, and Simulations," *Artificial Intelligence* 94 (1–2) (1997): 139–166.

[5] Barbara J. Grosz and Candace L. Sidner, "Plans for Discourse," in *Intentions in Communication*, ed. Philip R. Cohen, Jerry Morgan, and Martha E. Pollack (Cambridge, Mass.: MIT Press, 1990); and Milind Tambe, "Towards Flexible Teamwork," *Journal of Artificial Intelligence Research* 7 (1997): 83–124.

[6] Sarvapali D. Ramchurn, Trung Dong Huynh, Feng Wu, et al., "A Disaster Response System Based on Human-Agent Collectives," *Journal of Artificial Intelligence Research* 57 (2016): 661–708; and Matthew Johnson, Jeffrey Mark Bradshaw, Paul J. Feltovich, and Catholijn M. Jonker, "Coactive Design: Designing Support for Interdependence in Joint Activity," *Journal of Human-Robot Interaction* 3 (1) (2014): 43–69.

[7] Bryan Wilder, Eric Horvitz, and Ece Kamar, "Learning to Complement Humans," in *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, ed. Christian Bessiere (Santa Clara, Calif.: International Joint Conferences on Artificial Intelligence Organization, 2020), 1526–1533; and Amos Azaria, Zinovi Rabinovich, Sarit Kraus, et al., "Strategic Advice Provision in Repeated Human-Agent Interactions," *Autonomous Agents and Multi-Agent Systems* 30 (1) (2016): 4–29.

[8] Stefanos Nikolaidis, Przemyslaw Lasota, Ramya Ramakrishnan, and Julie Shah, "Improved Human-Robot Team Performance through Cross-Training, an Approach Inspired by Human Team Training Practices," *The International Journal of Robotics Research* 34 (14) (2015): 1711–1730.

[9] Iyad Rahwan, Manuel Cebrian, Nick Obradovich, and Josh Bongardet, "Machine Behaviour," *Nature* 568 (7753) (2019): 477–486; and Prashan Madumal, Tim Miller, Liz Sonenberg, and Frank Vetere, in "Explainable Reinforcement Learning through a Causal Lens," *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence* (Palo Alto, Calif.: AAAI Press, 2020).

[10] Amos Tversky and Daniel Kahneman, "Judgment under Uncertainty: Heuristics and Biases," *Science* 185 (4157) (1974): 1124–1131.

[11] Tathagata Chakraborti, Sarath Sreedharan, Yu Zhang, and Subbarao Kambhampati, "Plan Explanations as Model Reconciliation: Moving Beyond *Explanation as Soliloquy*," in *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17* (Santa Clara, Calif.: International Joint Conferences on Artificial Intelligence Organization, 2017), 156–163.

[12] Ece Kamar, Ya'akov Gal, and Barbara J. Grosz, "Incorporating Helpful Behavior into Collaborative Planning," in *Proceedings of the 8th International Conference on Autonomous Agents and Multiagent Systems*, ed. Ryszard Kowalczyk, Quoc Bao Vo, Zakaria Maamar, and Michael Huhns (Budapest: International Foundation for Autonomous Agents and Multiagent Systems, 2009); and Grosz and Kraus, "Collaborative Plans for Complex Group Action."

[13] Ofra Amir, Barbara J. Grosz, Krzysztof Z. Gajos, and Limor Gultchin, "Personalized Change Awareness: Reducing Information Overload in Loosely-Coupled Teamwork," *Artificial Intelligence* 275 (2019): 204–233.

[14] Nikolaidis et al., "Improved Human-Robot Team Performance."

[15] Ya'akov Gal, Barbara Grosz, Sarit Kraus, et al., "Agent Decision-Making in Open Mixed Networks," *Artificial Intelligence* 174 (18) (2010): 1460–1480.

[16] Azaria et al., "Strategic Advice Provision in Repeated Human Agent Interactions."

[17] Arlette van Wissen, Ya'akov Gal, Bart Kamphorst, and Virginia Dignum, "Human-Agent Teamwork in Dynamic Environments," *Computers in Human Behavior* 28 (1) (2012): 23–33.

[18] Ece Kamar, Ya'akov Kobi Gal, and Barbara J. Grosz, "Modeling Information Exchange Opportunities for Effective Human-Computer Teamwork," *Artificial Intelligence* 195 (2013).

[19] Amir et al., "Personalized Change Awareness."

[20] Avi Segal, Kobi Gal, Ece Kamar, et al., "Optimizing Interventions via Offline Policy Evaluation: Studies in Citizen Science," in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence* (Palo Alto, Calif.: AAAI Press, 2018).

[21] Ece Kamar, Severin Hacker, and Eric Horvitz, "Combining Human and Machine Intelligence in Large-Scale Crowdsourcing," in *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2012)* (Valencia, Spain: International Foundation for Autonomous Agents and Multiagent Systems, 2012); and Shengying Pan, Kate Larson, Josh Bradshaw, and Edith Law, "Dynamic Task Allocation Algorithm for Hiring Workers that Learn," in *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI-16*, ed. Subbarao Kambhampati (Santa Clara, Calif.: International Joint Conferences on Artificial Intelligence Organization, 2016).

[22] Avi Segal, Yossi Ben David, Joseph Jay Williams, et al., "Combining Difficulty Ranking with Multi-Armed Bandits to Sequence Educational Content," in *Artificial Intelligence in Education: 19th International Conference, AIED 2018, London, UK, June 27 – 30, 2018, Proceedings, Part II*, ed. Carolyn Penstein Rosé, Roberto Martínez-Maldonado, H. Ulrich Hoppe, et al. (New York: Springer, 2018), 317–321.

[23] Jonathan Martinez, Kobi Gal, Ece Kamar, and Levi H. S. Lelis, "Personalization in Human-AI Teams: Improving the Compatibility-Accuracy Tradeoff," in *Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence* (Palo Alto, Calif.: AAAI Press, 2021).

[24] Reuth Mirsky, Ya'akov Gal, and Stuart M. Shieber. "CRADLE: An Online Plan Recognition Algorithm for Exploratory Domains," *ACM Transactions on Intelligent Systems and Technology* 8 (3) (2017): 1–22; Chakraborti et al., "Plan Explanations as Model Reconciliation"; and Mor Vered, Gal A. Kaminka, and Sivan Biham, "Online Goal Recognition through Mirroring: Humans and Agents," presented at the Fourth Annual Conference on Advances in Cognitive Systems, Evanston, Illinois, June 23–26, 2016.

[25] Ofra Amir and Ya'akov Gal, "Plan Recognition and Visualization in Exploratory Learning Environments," *ACM Transactions on Interactive Intelligent Systems* 3 (3) (2013) : 1–23 ; and Nicholas Hoernle, Kobi Gal, Barbara Grosz, and Leilah Lyons, "Interpretable Models for Understanding Immersive Simulations," in *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence* (Santa Clara, Calif. : International Joint Conferences on Artificial Intelligence Organization, 2020).

[26] W. Bradley Knox and Peter Stone, "Combining Manual Feedback with Subsequent Reward Signals for Reinforcement Learning," in *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems*, vol. 1 (Toronto : International Foundation for Autonomous Agents and Multiagent Systems, 2010), 5–12.

[27] Travis Mandel, Yun-En Liu, Emma Brunskill, and Zoran Popovic, "Where to Add Actions in Human-in-the-Loop Reinforcement Learning," in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence* (Palo Alto, Calif. : AAAI Press, 2017).

[28] Matthew E. Taylor, Halit Bener Suay, and Sonia Chernova, "Integrating Reinforcement Learning with Human Demonstrations of Varying Ability," *Proceedings of the 10th International Conference on Autonomous Agents and Multiagent Systems* (Taipei : International Foundation for Autonomous Agents and Multiagent Systems, 2011).

[29] Avi Rosenfeld and Ariella Richardson, "Explainability in Human-Agent Systems," *Autonomous Agents and Multi-Agent Systems* 33 (3) (2019) ; and Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach, "Understanding the Effect of Accuracy on Trust in Machine Learning Models," in *Proceedings of the CHI Conference on Human Factors in Computing Systems* (New York : Association for Computing Machinery, 2019).

[30] Isaac Lage, Andrew Slavin Ross, Been Kim, et al., "Human-in-the-Loop Interpretability Prior," *Advances in Neural Information Processing Systems* 31 (2018) ; and Hoernle et al., "Interpretable Models for Understanding Immersive Simulations."

[31] Gal et al., "Agent Decision-Making in Open Mixed Networks" ; and van Wissen et al., "Human-Agent Teamwork in Dynamic Environments."

[32] Raz Lin, Sarit Kraus, Tim Baarslag, et al., "Genius : An Integrated Environment for Supporting the Design of Generic Automated Negotiators," *Computational Intelligence* 30 (1) (2014).

[33] Johnathan Mell, Gale Lucas, Sharon Mozgai, and Jonathan Gratch, "The Effects of Experience on Deception in Human-Agent Negotiation," *Journal of Artificial Intelligence Research* 68 (2020) : 633–660.

[34] See Automated Negotiating Agents Competition (ANAC), http ://ii.tudelft.nl/nego/node/7 (accessed March 1, 2022).

[35] Ofra Amir, Barbara J. Grosz, Krzysztof Z. Gajos, et al., "From Care Plans to Care Coordination : Opportunities for Computer Support of Teamwork in Complex Healthcare," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (New York : Association for Computing Machinery, 2015), 1419–1428 ; and Mary L. Gray and Siddharth Suri, *Ghost Work : How to Stop Silicon Valley from Building a New Global Underclass* (New York : Eamon Dolan Books, 2019).

[36] Mell et al., "The Effects of Experience on Deception in Human-Agent Negotiation" ; and Azaria, "Strategic Advice Provision in Repeated Human Agent Interactions."

Human Language Understanding & Reasoning

Christopher D. Manning

The last decade has yielded dramatic and quite surprising breakthroughs in natural language processing through the use of simple artificial neural network computations, replicated on a very large scale and trained over exceedingly large amounts of data. The resulting pretrained language models, such as BERT and GPT-3, have provided a powerful universal language understanding and generation base, which can easily be adapted to many understanding, writing, and reasoning tasks. These models show the first inklings of a more general form of artificial intelligence, which may lead to powerful foundation models in domains of sensory experience beyond just language.

When scientists consider artificial intelligence, they mostly think of modeling or recreating the capabilities of an individual human brain. But modern human intelligence is much more than the intelligence of an individual brain. Human language is powerful and has been transformative to our species because it gives groups of people a way to network human brains together. An individual human may not be much more intelligent than our close relatives of chimpanzees or bonobos. These apes have been shown to possess many of the hallmark skills of human intelligence, such as using tools and planning; moreover, they have better short-term memory than we do.[1] When humans invented language is still, and perhaps will forever be, quite uncertain, but within the long evolutionary history of life on Earth, human beings developed language incredibly recently. The common ancestor of prosimians, monkeys, and apes dates to perhaps sixty-five million years ago; humans separated from chimps perhaps six million years ago, while human language is generally assumed to be only a few hundred thousand years old.[2] Once humans developed language, the power of communication quickly led to the ascendancy of Homo sapiens over other creatures, even though we are not as strong as an elephant nor as fast as a cheetah. It was much more recently that humans developed writing (only a bit more than five thousand years ago), allowing knowledge to be communicated across distances of time and space. In just a few thousand years, this information-sharing mechanism took us from the bronze age to the smartphones of today. A high-fidelity

© 2022 by Christopher D. Manning
Published under a Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) license
https://doi.org/10.1162/DAED_a_01905

code allowing both rational discussion among humans and the distribution of information has allowed the cultural evolution of complex societies and the knowledge underlying modern technologies. The power of language is fundamental to human societal intelligence, and language will retain an important role in a future world in which human abilities are augmented by artificial intelligence tools.

For these reasons, the field of natural language processing (NLP) emerged in tandem with the earliest developments in artificial intelligence. Indeed, initial work on the NLP problem of machine translation, including the famous Georgetown-IBM demonstration in 1954, slightly preceded the coining of the term "artificial intelligence" in 1956.[3] In this essay, I give a brief outline of the history of natural language processing. I then describe the dramatic recent developments in NLP coming from the use of large artificial neural network models trained on very large amounts of data. I trace the dramatic progress that has been made in building effective NLP systems using these techniques, and conclude with some thoughts on what these models achieve and where things will head next.

The history of natural language processing until now can be roughly divided into four eras. The first era runs from 1950 to 1969. NLP research began as research in machine translation. It was imagined that translation could quickly build on the great successes of computers in code breaking during World War II. On both sides of the Cold War, researchers sought to develop systems capable of translating the scientific output of other nations. Yet, at the beginning of this era, almost nothing was known about the structure of human language, artificial intelligence, or machine learning. The amount of computation and data available was, in retrospect, comically small. Although initial systems were promoted with great fanfare, the systems provided little more than word-level translation lookups and some simple, not very principled rule-based mechanisms to deal with the inflectional forms of words (morphology) and word order.

The second era, from 1970 to 1992, saw the development of a whole series of NLP demonstration systems that showed sophistication and depth in handling phenomena like syntax and reference in human languages. These systems included SHRDLU by Terry Winograd, LUNAR by Bill Woods, Roger Schank's systems such as SAM, Gary Hendrix's LIFER, and GUS by Danny Bobrow.[4] These were all hand-built, rule-based systems, but they started to model and use some of the complexity of human language understanding. Some systems were even deployed operationally for tasks like database querying.[5] Linguistics and knowledge-based artificial intelligence were rapidly developing, and in the second decade of this era, a new generation of hand-built systems emerged, which had a clear separation between declarative linguistic knowledge and its procedural processing, and which benefited from the development of a range of more modern linguistic theories.

However, the direction of work changed markedly in the third era, from roughly 1993 to 2012. In this period, digital text became abundantly available, and the compelling direction was to develop algorithms that could achieve some level of language understanding over large amounts of natural text and that used the existence of this text to help provide this ability. This led to a fundamental reorientation of the field around empirical machine learning models of NLP, an orientation that still dominates the field today. At the beginning of this period, the dominant modus operandi was to get hold of a reasonable quantity of online text – in those days, text collections were generally in the low tens of millions of words – and to extract some kind of model out of these data, largely by counting particular facts. For example, you might learn that the kinds of things people *capture* are fairly evenly balanced between locations with people (like a *city*, *town*, or *fort*) and metaphorical notions (like *imagination*, *attention*, or *essence*). But counts on words only go so far in providing language understanding devices, and early empirical attempts to learn language structure from text collections were fairly unsuccessful.[6] This led most of the field to concentrate on constructing annotated linguistic resources, such as labeling the sense of words, instances of person or company names in texts, or the grammatical structure of sentences in treebanks, followed by the use of supervised machine learning techniques to build models that could produce similar labels on new pieces of text at runtime.

The period from 2013 to present extended the empirical orientation of the third era, but the work has been enormously changed by the introduction of deep learning or artificial neural network methods. In this approach, words and sentences are represented by a position in a (several hundred- or thousand-dimensional) real-valued vector space, and similarities of meaning or syntax are represented by proximity in this space. From 2013 to 2018, deep learning provided a more powerful method for building performant models: it was easier to model longer distance contexts, and models generalized better to words or phrases with similar meanings because they could exploit proximity in a vector space, rather than depending on the identity of symbols (such as word form or part of speech). Nevertheless, the approach was unchanged in building supervised machine learning models to perform particular analysis tasks. Everything changed in 2018, when NLP was the first major success of very large scale *self-supervised* neural network learning. In this approach, systems can learn an enormous amount of knowledge of a language and the world simply from being exposed to an extremely large quantity of text (now normally in the billions of words). The method of self-supervision by which this is done is for the system to create from the text its own prediction challenges, such as successively identifying each next word in the text given the previous words or filling in a masked word or phrase in a text. By repeating such prediction tasks billions of times and learning from its mistakes, so the model does better next time given a similar textual context, general knowledge of a language and the world is

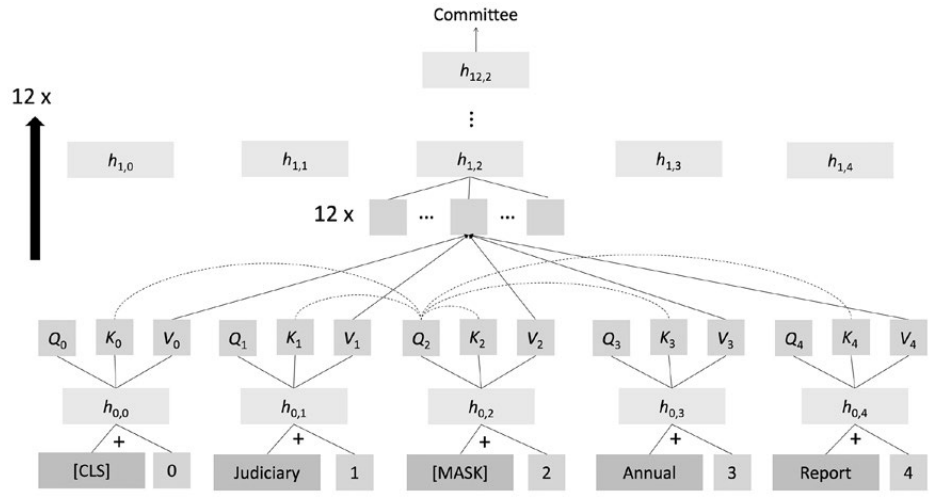
accumulated, and this knowledge can then be deployed for tasks of interest, such as question answering or text classification.

In hindsight, the development of large-scale self-supervised learning approaches may well be viewed as the fundamental change, and the third era might be extended until 2017. The impact of pretrained self-supervised approaches has been revolutionary: it is now possible to train models on huge amounts of unlabeled human language material in such a way as to produce one large pretrained model that can be very easily adapted, via fine-tuning or prompting, to give strong results on all sorts of natural language understanding and generation tasks. As a result, progress and interest in NLP have exploded. There is a sense of optimism that we are starting to see the emergence of knowledge-imbued systems that have a degree of general intelligence.

I cannot give here a full description of the now-dominant neural network models of human language, but I can offer an inkling. These models represent everything via vectors of real numbers and are able to learn good representations after exposure to many pieces of data by back-propagation of errors (which comes down to doing differential calculus) from some prediction task back to the representations of the words in a text. Since 2018, the dominant neural network model for NLP applications has been the transformer neural network.[7] With several ideas and parts, a transformer is a much more complex model than the simple neural networks for sequences of words that were explored in earlier decades. The dominant idea is one of *attention*, by which a representation at a position is computed as a weighted combination of representations from other positions. A common self-supervision objective in a transformer model is to mask out occasional words in a text. The model works out what word used to be there. It does this by calculating from each word position (including mask positions) vectors that represent a query, key, and value at that position. The query at a position is compared with the value at every position to calculate how much attention to pay to each position; based on this, a weighted average of the values at all positions is calculated. This operation is repeated many times at each level of the transformer neural net, and the resulting value is further manipulated through a fully connected neural net layer and through use of normalization layers and residual connections to produce a new vector for each word. This whole process is repeated many times, giving extra layers of depth to the transformer neural net. At the end, the representation above a mask position should capture the word that was there in the original text: for instance, *committee* as illustrated in Figure 1.

It is not at all obvious what can be achieved or learned by the many simple calculations of a transformer neural net. At first, this may sound like some kind of complex statistical association learner. However, given a very powerful, flexible, and high-parameter model like a transformer neural net and an enormous amount

Figure 1
Details of the Attention Calculations in One Part of a Transformer Neural Net Model



From this calculation, the transformer neural net is able to predict the word *committee* in the masked position.

of data to practice predictions on, these models discover and represent much of the structure of human languages. Indeed, work has shown that these models learn and represent the syntactic structure of a sentence and will learn to memorize many facts of the world, since each of these things helps the model to predict masked words successfully.[8] Moreover, while predicting a masked word initially seems a rather simple and low-level task – a kind of humorless Mad Libs – and not something sophisticated, like diagramming a sentence to show its grammatical structure, this task turns out to be very powerful because it is universal: every form of linguistic and world knowledge, from sentence structure, word connotations, and facts about the world, help one to do this task better. As a result, these models assemble a broad general knowledge of the language and world to which they are exposed. A single such large pretrained language model (LPLM) can be deployed for many particular NLP tasks with only a small amount of further instruction. The standard way of doing this from 2018 to 2020 was fine-tuning the model via a small amount of additional supervised learning, training it on the exact task of interest. But very recently, researchers have surprisingly found that the largest of these models, such as GPT-3 (Generative Pre-trained Transformer-3),

can perform novel tasks very well with just a *prompt*. Give them a human language description or several examples of what one wants them to do, and they can perform many tasks for which they were never otherwise trained.[9]

Traditional natural language processing models were elaborately composed from several usually independently developed components, frequently built into a pipeline, which first tried to capture the sentence structure and low-level entities of a text and then something of the higher-level meaning, which would be fed into some domain-specific execution component. In the last few years, companies have started to replace such traditional NLP solutions with LPLMs, usually fine-tuned to perform particular tasks. What can we expect these systems to do in the 2020s?

Early machine translation systems covered limited linguistic constructions in a limited domain.[10] Building large statistical models from parallel corpora of translated text made broad-coverage machine translation possible, something that most people first experienced using Google Translate after it launched in 2006. A decade later, in late 2016, Google's machine translation improved markedly when they switched to the use of neural machine translation.[11] But that system had a shorter lifespan: transformer-based neural translation was rolled out in 2020.[12] This new system improved not only via a different neural architecture but via use of a fundamentally different approach. Rather than building numerous pairwise systems from parallel text that translate between two languages, the new system gains from one huge neural net that was simultaneously trained on all languages that Google Translate covers, with input simply marked by a distinct token that indicates the language. While this system still makes mistakes and machine translation research continues, the quality of automatic translation today is remarkable. When I enter a couple of sentences from today's *Le Monde* culture section:

> *Il avait été surnommé, au milieu des années 1930, le « Fou chantant », alors qu'il faisait ses débuts d'artiste soliste après avoir créé, en 1933, un duo à succès avec le pianiste Johnny Hess. Pour son dynamisme sur scène, silhouette agile, ses yeux écarquillés et rieurs, ses cheveux en bataille, surtout pour le rythme qu'il donnait aux mots dans ses interprétations et l'écriture de ses textes.*[13]

the translation is excellent:

> *He was nicknamed the Singing Madman in the mid-1930s when he was making his debut as a solo artist after creating a successful duet with pianist Johnny Hess in 1933. For his dynamism on stage, his agile figure, his wide, laughing eyes, his messy hair, especially for the rhythm he gave to the words in his interpretations and the writing of his texts.*

In *question answering*, a system looks for relevant information across a collection of texts and then provides answers to specific questions (rather than just re-

turning pages that are suggested to hold relevant information, as in the early generations of Web search). Question answering has many straightforward commercial applications, including both presale and postsale customer support. Modern neural network question-answering systems have high accuracy in extracting an answer present in a text and are even fairly good at working out that no answer is present. For example, from this passage:

> Samsung saved its best features for the Galaxy Note 20 Ultra, including a more refined design than the Galaxy S20 Ultra – a phone I don't recommend. You'll find an exceptional 6.9-inch screen, sharp 5x optical zoom camera and a swifter stylus for annotating screenshots and taking notes. The Note 20 Ultra also makes small but significant enhancements over the Note 10 Plus, especially in the camera realm. Do these features justify the Note 20 Ultra's price? It begins at $1,300 for the 128GB version. The retail price is a steep ask, especially when you combine a climate of deep global recession and mounting unemployment.

One can get answers to questions like the following (using the UnifiedQA model):[14]

> How expensive is the Samsung Galaxy Note 20 Ultra?
> *$1,300 for the 128GB version*
>
> Does the Galaxy Note 20 Ultra have 20x optical zoom?
> *no*
>
> What is the optical zoom of the Galaxy Note 20 Ultra?
> *5x*
>
> How big is the screen of the Galaxy Note 20 Ultra?
> *6.9-inch*

For common traditional NLP tasks like marking person or organization names in a piece of text or classifying the sentiment of a text about a product (as positive or negative), the best current systems are again based on LPLMs, usually fine-tuned by providing a set of examples labeled in the desired way. While these tasks could be done quite well even before recent large language models, the greater breadth of knowledge of language and the world in these models has further improved performance on these tasks.

Finally, LPLMs have led to a revolution in the ability to generate fluent and connected text. In addition to many creative uses, such systems have prosaic uses ranging from writing formulaic news articles like earnings or sports reports and automating summarization. For example, such a system can help a radiologist by suggesting the impression (or summary) based on the radiologist's findings. For the findings below, we can see that the system-generated impression is quite similar to a radiologist-generated impression:[15]

> Findings: lines/tubes: right ij sheath with central venous catheter tip overlying the svc. on initial radiograph, endotracheal tube between the clavicular heads, and enteric tube with side port at the ge junction and tip below the diaphragm off the field-of-view; these are removed on subsequent film. mediastinal drains and left thoracostomy tube are unchanged. lungs: low lung volumes. retrocardiac airspace disease, slightly increased on most recent film. pleura: small left pleural effusion. no pneumothorax. heart and mediastinum: postsurgical widening of the cardiomediastinal silhouette. aortic arch calcification. bones: intact median sternotomy wires.
>
> Radiologist-generated impression: left basilar airspace disease and small left pleural effusion. lines and tubes positioned as above.
>
> System-generated impression: lines and tubes as described above. retrocardiac airspace disease, slightly increased on most recent film. small left pleural effusion.

These recent NLP systems perform very well on many tasks. Indeed, given a fixed task, they can often be trained to perform it as well as human beings, on average. Nevertheless, there are still reasons to be skeptical as to whether these systems really understand what they are doing, or whether they are just very elaborate rewriting systems, bereft of meaning.

The dominant approach to describing meaning, in not only linguistics and philosophy of language but also for programming languages, is a *denotational semantics* approach or a *theory of reference*: the meaning of a word, phrase, or sentence is the set of objects or situations in the world that it describes (or a mathematical abstraction thereof). This contrasts with the simple *distributional semantics* (or *use theory of meaning*) of modern empirical work in NLP, whereby the meaning of a word is simply a description of the contexts in which it appears.[16] Some have suggested that the latter is not a theory of semantics at all but just a regurgitation of distributional or syntactic facts.[17] I would disagree. Meaning is not all or nothing; in many circumstances, we partially appreciate the meaning of a linguistic form. I suggest that meaning arises from understanding the network of connections between a linguistic form and other things, whether they be objects in the world or other linguistic forms. If we possess a dense network of connections, then we have a good sense of the meaning of the linguistic form. For example, if I have held an Indian *shehnai*, then I have a reasonable idea of the meaning of the word, but I would have a richer meaning if I had also heard one being played. Going in the other direction, if I have never seen, felt, or heard a *shehnai*, but someone tells me that *it's like a traditional Indian oboe*, then the word has some meaning for me: it has connections to India, to wind instruments that use reeds, and to playing music. If someone added that *it has holes sort of like a recorder, but it has multiple reeds and a flared end more like an oboe*, then I have more network con-

nections to objects and attributes. Conversely, I might not have that information but just a couple of contexts in which the word has been used, such as: *From a week before, shehnai players sat in bamboo machans at the entrance to the house, playing their pipes. Bikash Babu disliked the shehnai's wail, but was determined to fulfil every conventional expectation the groom's family might have*.[18] Then, in some ways, I understand the meaning of the word *shehnai* rather less, but I still know that it is a pipe-like musical instrument, and my meaning is not a subset of the meaning of the person who has simply held a *shehnai*, for I know some additional cultural connections of the word that they lack.

Using this definition whereby understanding meaning consists of understanding networks of connections of linguistic forms, there can be no doubt that pretrained language models learn meanings. As well as word meanings, they learn much about the world. If they are trained on encyclopedic texts (as they usually are), they will learn that Abraham Lincoln was born in 1809 in Kentucky and that the lead singer of Destiny's Child was Beyoncé Knowles-Carter. Our machines can richly benefit from writing as a store of human knowledge, just like people. Nevertheless, the models' word meanings and knowledge of the world are often very incomplete and cry out for being augmented with other sensory data and knowledge. Large amounts of text data provided a very accessible way first to explore and build these models, but it will be useful to expand to other kinds of data.

The success of LPLMs on language-understanding tasks and the exciting prospects for extending large-scale self-supervised learning to other data modalities – such as vision, robotics, knowledge graphs, bioinformatics, and multimodal data – suggests exploring a more general direction. We have proposed the term *foundation models* for the general class of models with millions of parameters trained on copious data via self-supervision that can then easily be adapted to perform a wide range of downstream tasks.[19] LPLMs like BERT (Bidirectional Encoder Representations from Transformers) and GPT-3 are early examples of foundation models, but work is now underway more broadly.[20] One direction is to connect language models with more structured stores of knowledge represented as a knowledge graph neural network or as a large supply of text to be consulted at runtime.[21] However, the most exciting and promising direction is to build foundation models that also take in other sensory data from the world to enable integrated, multimodal learning. An example of this is the recent DALL·E model that, after self-supervised learning on a corpus of paired images and text, can express the meaning of a new piece of text by producing a corresponding picture.[22]

We are still very early in the era of foundation models, but let me sketch a possible future. Most information processing and analysis tasks, and perhaps even things like robotic control, will be handled by a specialization of one of a relatively small number of foundation models. These models will be expensive and time-consuming to train, but adapting them to different tasks will be quite easy;

indeed, one might be able to do it simply with natural language instructions. This resulting convergence on a small number of models carries several risks: the groups capable of building these models may have excessive power and influence, many end users might suffer from any biases present in these models, and it will be difficult to tell if models are safe to use in particular contexts because the models and their training data are so large. Nevertheless, the ability of these models to deploy knowledge gained from a huge amount of training data to many different runtime tasks will make these models powerful, and they will for the first time demonstrate the artificial intelligence goal of one machine learning model doing many particular tasks based on simply being instructed on the spot as to what it should do. While the eventual possibilities for these models are only dimly understood, they are likely to remain limited, lacking a human-level ability for careful logical or causal reasoning. But the broad effectiveness of foundation models means that they will be very widely deployed, and they will give people in the coming decade their first glimpses of a more general form of artificial intelligence.

ABOUT THE AUTHOR

Christopher D. Manning is the Thomas M. Siebel Professor in Machine Learning and Professor of Linguistics and of Computer Science at Stanford University; Director of the Stanford Artificial Intelligence Laboratory (SAIL); and Associate Director of the Stanford Institute for Human-Centered Artificial Intelligence (HAI). He also served as President of the Association for Computational Linguistics. He is the author of *Introduction to Information Retrieval* (with Hinrich Schütze and Prabhakar Raghavan, 2008), *Foundations of Statistical Natural Language Processing* (with Hinrich Schütze, 1999), and *Complex Predicates and Information Spreading in LFG* (with Avery Andrews, 1999).

ENDNOTES

[1] Frans de Waal, *Are We Smart Enough to Know How Smart Animals Are?* (New York: W. W. Norton, 2017).

[2] Mark Pagel, "Q&A: What Is Human Language, When Did It Evolve and Why Should We Care?" *BMC Biology* 15 (1) (2017): 64.

[3] W. John Hutchins, "The Georgetown-IBM Experiment Demonstrated in January 1954," in *Machine Translation: From Real Users to Research*, ed. Robert E. Frederking and Kathryn B. Taylor (New York: Springer, 2004), 102–114.

[4] A survey of these systems and references to individual systems appears in Avron Barr, "Natural Language Understanding," *AI Magazine*, Fall 1980.

[5] Larry R. Harris, "Experience with Robot in 12 Commercial, Natural Language Data Base Query Applications" in *Proceedings of the 6th International Joint Conference on Artificial Intelligence, IJCAI-79* (Santa Clara, Calif.: International Joint Conferences on Artificial Intelligence Organization, 1979), 365–371.

[6] Glenn Carroll and Eugene Charniak, "Two Experiments on Learning Probabilistic Dependency Grammars from Corpora," in *Working Notes of the Workshop Statistically-Based NLP Techniques*, ed. Carl Weir, Stephen Abney, Ralph Grishman, and Ralph Weischedel (Menlo Park, Calif.: AAAI Press, 1992).

[7] Ashish Vaswani, Noam Shazeer, Niki Parmar, et al., "Attention Is All You Need," *Advances in Neural Information Processing Systems* 30 (2017).

[8] Christopher D. Manning, Kevin Clark, John Hewitt, et al., "Emergent Linguistic Structure in Artificial Neural Networks Trained by Self-Supervision," *Proceedings of the National Academy of Sciences* 117 (48) (2020): 30046–30054.

[9] Tom Brown, Benjamin Mann, Nick Ryder, et al., "Language Models Are Few-Shot Learners," *Advances in Neural Information Processing Systems* 33 (2020): 1877–1901.

[10] For example, Météo translated Canadian weather reports between French and English; see Monique Chevalier, Jules Dansereau, and Guy Poulin, *TAUM-MÉTÉO: Description du système* (Montreal: Traduction Automatique à l'Université de Montréal, 1978).

[11] Gideon Lewis-Kraus, "The Great A.I. Awakening," *The New York Times Magazine*, December 18, 2016.

[12] Isaac Caswell and Bowen Liang, "Recent Advances in Google Translate," Google AI Blog, June 8, 2020, https://ai.googleblog.com/2020/06/recent-advances-in-google-translate.html.

[13] Sylvain Siclier, "A Paris, le Hall de la chanson fête les inventions de Charles Trenet," *Le Monde*, June 16, 2021, https://www.lemonde.fr/culture/article/2021/06/16/a-paris-le-hall-de-la-chanson-fete-les-inventions-de-charles-trenet_6084391_3246.html.

[14] Daniel Khashabi, Sewon Min, Tushar Khot, et al., "UnifiedQA: Crossing Format Boundaries with a Single QA System," in *Findings of the Association for Computational Linguistics: EMNLP 2020* (Stroudsburg, Pa.: Association for Computational Linguistics, 2020), 1896–1907.

[15] Yuhao Zhang, Derek Merck, Emily Bao Tsai, et al., "Optimizing the Factual Correctness of a Summary: A Study of Summarizing Radiology Reports," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (Stroudsburg, Pa.: Association for Computational Linguistics, 2020), 5108–5120.

[16] For an introduction to this contrast, see Gemma Boleda and Aurélie Herbelot, "Formal Distributional Semantics: Introduction to the Special Issue," *Computational Linguistics* 42 (4) (2016): 619–635.

[17] Emily M. Bender and Alexander Koller, "Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (Stroudsburg, Pa.: Association for Computational Linguistics, 2020), 5185–5198.

[18] From Anuradha Roy, *An Atlas of Impossible Longing* (New York: Free Press, 2011).

[19] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, et al., "On the Opportunities and Risks of Foundation Models," arXiv (2021), https://arxiv.org/abs/2108.07258.

[20] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics* (Stroudsburg, Pa.: Association for Computational Linguistics, 2019), 4171–4186.

[21] Robert Logan, Nelson F. Liu, Matthew E. Peters, et al., "Barack's Wife Hillary: Using Knowledge Graphs for Fact-Aware Language Modeling," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (Stroudsburg, Pa.: Association for Computational Linguistics, 2019), 5962–5971; and Kelvin Guu, Kenton Lee, Zora Tung, et al., "REALM: Retrieval-Augmented Language Model Pre-Training," *Proceedings of Machine Learning Research* 119 (2020).

[22] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, et al., "Zero-Shot Text-To-Image Generation," arXiv (2021), https://arxiv.org/abs/2102.12092.

The Curious Case of Commonsense Intelligence

Yejin Choi

Commonsense intelligence is a long-standing puzzle in AI. Despite considerable advances in deep learning, AI continues to be narrow and brittle due to its lack of common sense. Why is common sense so trivial for humans but so hard for machines? In this essay, I map the twists and turns in recent research adventures toward commonsense AI. As we will see, the latest advances on common sense are riddled with new, potentially counterintuitive perspectives and questions. In particular, I discuss the significance of language for modeling intuitive reasoning, the fundamental limitations of logic formalisms despite their intellectual appeal, the case for on-the-fly generative reasoning through language, the continuum between knowledge and reasoning, and the blend between symbolic and neural knowledge representations.

Commonsense intelligence is a long-standing challenge in AI. Despite considerable advances in deep learning, AI systems continue to be narrow and brittle. One of the fundamental limitations of AI can be characterized as its lack of commonsense intelligence: the ability to reason intuitively about everyday situations and events, which requires rich background knowledge about how the physical and social world works.[1]

Trivial for humans, acquiring commonsense intelligence has been considered a nearly impossible goal in AI. In fact, until several years ago, the word "commonsense" was considered taboo for anyone wanting to be taken seriously in the mainstream research community. How, then, is this goal *now* feasible? To help answer this question, we will characterize what approaches have been tried in the past and what alternative paths have yet to be explored.

First and foremost, the significance of language – not just words and phrases, but the full scope of natural language – has long been overlooked as a representation medium for modeling commonsense knowledge and reasoning. At first glance, language seems too imprecise and variable, thus, many earlier efforts sought logic-based formalisms to describe commonsense rules for machines. But despite their intellectual appeal, logic-based formalisms proved too brittle to scale beyond experimental toy problems. In contrast, *language-based formalisms*, despite their apparent imprecision and variability, are sufficiently expressive and robust to encom-

© 2022 by Yejin Choi
Published under a Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) license
https://doi.org/10.1162/DAED_a_01906

pass the vast number of commonsense facts and rules about how the world works. After all, it is language, not logical forms, through which humans acquire knowledge about the world. And this holds true despite the ambiguities of language and the inconsistencies of knowledge reported in books, news, and even the scientific literature. Thus, in order to match the scale and complexity of human-level knowledge acquisition, AI cannot go far without direct integration of language.

Second, most prior efforts were developed in the pre–deep learning era, without benefiting from large-scale data, compute, and neural networks. Deep learning presents entirely new opportunities for training neural commonsense models using a massive amount of raw text, fused with symbolic commonsense knowledge graphs. Again, the switch to language-based formalisms is the key to benefit from the empirical breakthroughs of deep neural networks, as it allows for powerful transfer learning from *language* models to *knowledge* models.

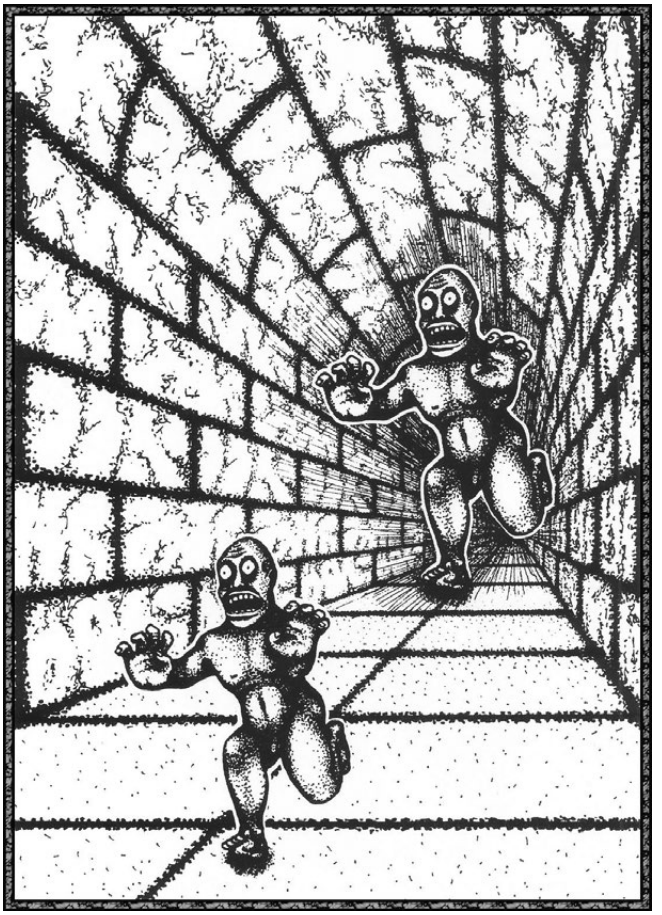
The landscape has changed considerably over the past few years. The Allen Institute for Artificial Intelligence created the research project Mosaic, which I lead, to focus on commonsense intelligence.[2] The Association for Computational Linguistics (ACL), which hosts one of the premiere conferences in AI focusing on human language technologies, featured a tutorial on commonsense knowledge that attracted a great deal of attention from the research community.[3] Defense Advanced Research Projects Agency (DARPA), an agency of the U.S. Department of Defense, has also launched the Machine Common Sense (MCS) program to accelerate research on commonsense AI.[4]

Experience thus far suggests that research toward commonsense AI requires rethinking and challenging some of the most fundamental assumptions in the current paradigms of machine learning and AI. It also challenges our conceptual understanding about knowledge, reasoning, and language. As a result, it is inevitable that the perspectives discussed in this essay can appear counterintuitive or even controversial. As a starting point, let us examine intuitive reasoning and its connection to language generation.

Intuitive reasoning is effortless. Humans do it all the time, subconsciously, about nearly every object, person, and event that we encounter in our every waking moment. It is intuitive reasoning through which we make snap judgments about the big picture context of a scene that we observe only in part: the likely causes and effects of events, what might have happened before and what might happen next, what might be the motivations and intents of people, and what might be their mental and emotional states. Because intuitive reasoning is so natural and effortless, it is tempting to assume that it must be easy for AI as well.

A concrete example offers insight into why AI in the current paradigm might be far from reaching human-level intuitive reasoning on trivial everyday events and scenes. Consider psychologist Roger Shepard's optical illusion *Terror Subterra*,

Figure 1
Roger Shepard's *Terror Subterra*



Source: Roger Shepard, "Terror Subterra," in *Mind Sights: Original Visual Illusions, Ambiguities, and other Anomalies* (New York: W. H. Freeman & Co, 1990).

shown as Figure 1.[5] State-of-the-art computer vision systems are now capable of correctly identifying the literal content of the visual scene, such as objects and locations; in this case, two monsters in a tunnel. However, human-level cognitive understanding of the visual scene requires seeing beyond pixels: reasoning about the whole dynamic story that goes beyond the static scene captured in a still image. For example, we reason that the monsters are running, one is chasing another, and the chaser has hostile intentions while the chased is afraid.

This example leads us to unpack several interconnected insights: 1) intuitive reasoning is generative and instantaneous (as opposed to thoroughly discriminative across all possible alternatives); 2) the space of such reasoning is infinite, and thus requires the full scope of natural language to describe them (as opposed to a fixed set of predefined labels to choose from); 3) intuitive inferences are predictive in nature, and are therefore almost always defeasible with additional context; and 4) intuitive inferences draw from rich background knowledge about how the physical and social world works (as will be elaborated below).

What is remarkable about intuitive reasoning is that we make all these inferences instantaneously without ever enumerating and weighing all the other plausible but less likely, or implausible, inferences. For example, we do not consider plausible but less likely inferences about our monsters in the tunnel, like the monsters are running backward or are standing still on one foot. Nor do we consider outright implausible inferences, like the monsters are lying down on the floor or swimming in the ocean. Such less plausible or outright implausible inferences do not even come to our conscious mind. In fact, coming up with less likely or implausible alternatives can be effortful.

In other words, when we communicate our intuitive inferences in language, it is almost as if we generate the most likely intuitive inferences on the fly, word by word, without explicitly acknowledging the alternatives. This is analogous to how we can "think out loud": we can speak out the next word of a thought without first internally finishing the rest of the thought or planning the exact wordings of the sentences to come.

This is in stark contrast with how machine learning benchmarks – especially reasoning tasks – are most commonly formulated: as categorization tasks over a fixed set of predefined labels. Under such discriminative task formulations, models need to go through all possible labels one by one and choose the label with the highest score. Discriminative task formulations are effective for relatively narrowly defined tasks, such as object categorization in an image. However, human-level intuitive inferences require complex compositional reasoning over diverse concepts, including objects, actions, locations, attributes, and emotions. In other words, the space of concepts is infinite, as concepts can be composed of other concepts recursively. This is a point also emphasized by cognitive scientist Douglas Hofstadter and psychologist Emmanuel Sander in their book *Surfaces and Essences*: the set of concepts vastly outnumbers the set of words, and many concepts require open-text descriptions for lack of existing words or fixed phrases.[6]

This compositional nature of intuitive inferences has two important implications. First, natural language, not just words or phrases but the full scope of open-text descriptions, is the best way to communicate the content of intuitive inferences between humans and machines. Inventing a new labeling scheme (or logic

formalisms) can only be error prone and incomplete, since there always will be a significant representation gap between the labeling scheme and natural language. Second, the total number of all possible textual descriptions of intuitive inferences is too large for us, and even for AI, to enumerate and examine one by one in real time.

These observations motivate the need for computational models that can handle on-the-fly generative reasoning through language. The key underlying challenge is *scale*. Naively increasing the set of labels for discriminative models will not scale effectively to handle the sheer scope of intuitive reasoning, which requires complex and potentially novel compositional reasoning over diverse concepts. This calls for new machine-learning models and algorithms that can learn to generate intuitive inferences on the fly, word by word, just like how humans communicate their thoughts.

In fact, such word-by-word generation is exactly how text generation from neural language models operates today. For example, OpenAI's GPT-3 (Generative Pre-trained Transformer 3) – a language model that uses deep learning to produce speech-like text – has generated remarkably coherent paragraphs by sampling just one word at a time, without explicitly enumerating all other alternative sentences.[7] Advances in neural language models provide strong technical foundations to build language-based on-the-fly generative reasoning systems. Promising recent research is based on such generative reasoning: abductive reasoning, counterfactual story revision, and commonsense reasoning. But before we get there, let us discuss the importance of defeasible reasoning and commonsense knowledge.

When we look at Roger Shepard's monsters in a tunnel, it is reasonable to infer that one monster is chasing another, with emotions to match. But the faces of the two monsters are in fact identical: it is our brain projecting a story onto the image to the point of hallucinating two faces expressing visually distinct emotions. This story projection comes from our prior knowledge about how the world works, that when a monster is chasing, it is likely to have a hostile intent, while the chased would likely feel scared. Yet none of these is absolutely true and all can be defeated with additional context. For example, if we learned that these particular monsters have kind hearts despite their appearances, or that they are in fact practicing a new dance move, we would revise what we infer about their likely intents, emotions, and mental states.

Intuitive inferences draw from the rich background knowledge about how the world works, ranging from native physics to folk psychology. In order to close the gap between AI and humans in their intuitive reasoning capabilities over diverse everyday scenes and events, we need deep integration of language, and we need broad-coverage commonsense models of how the physical and the social world works.

Why does formal logic fail to model human reasoning? In their book *The Enigma of Reason*, cognitive scientists Hugo Mercier and Dan Sperber argue that "Reason is a mechanism of intuitive inferences . . . in which logic plays at best a marginal role."[8] Yet a dominant perspective underlying AI research is that human reasoning is modeled through a formal logic framework. The intellectual appeal of formal logic is its emphasis on correctness, a property that seems hard to dispute in itself. What could possibly go wrong with being correct?

There are two related challenges: the *purpose* and the *scale* of reasoning. The purpose of intuitive reasoning is to anticipate and predict what might be plausible explanations for our partial observations, so we can read between the lines in text and see beyond the frame of the image. As we have discussed, this means intuitive reasoning is almost always defeasible with additional context. Therefore, a reasoning framework that only seeks truthful conclusions is off point since it would rarely generate the sorts of rich conclusions that intuitive reasoning does.

The bigger challenge is the scale or the scope of reasoning. The reasoning framework, to be practically useful, should be ready to cover the full spectrum of concepts and compositions of concepts that we encounter in our everyday physical and social interactions with the world. In addition, the real world is filled with previously unseen situations, which require creative generation of hypotheses, novel compositions of concepts, and novel discovery of reasoning rules. In contrast, formal logic almost always assumes that some oracle will provide a predefined set of logic variables and logic implication rules. There is no such oracle. To date, we do not yet know how to automatically populate such logical representations of concepts and implication rules at scale, and those manually constructed by scientists have proven to be, time and again, too narrow in scope and too brittle to generalize. Moreover, formal logic frameworks fall short of providing practical solutions to the creative generation of hypotheses, novel compositions of concepts, and novel discovery of reasoning rules.

In regard to the defeasibility of intuitive reasoning, one might wonder whether adding probability models on top of formal logic frameworks could trivially address this challenge, since probabilistic logic frameworks can generate uncertain conclusions that are defeasible. The real bottleneck of scale is not due to lack of probabilistic measures of uncertainty, however. Adding probabilistic models over a small, fixed set of variables and logical rules does not automatically increase the diversity and complexity of concepts covered by the logical forms. The challenge of automatically populating formal logical variables and implication rules still remains, with or without probabilistic measures on top.

Logical reasoning is often associated with deductive reasoning and inductive reasoning. Deduction starts with a general rule, which is then applied to a concrete case, whereas induction begins with facts about individual

cases, which are then generalized to a general rule. But the scope of deduction and induction together is only the tip of the iceberg of human reasoning. Indeed, neither deduction nor induction can account for the sorts of intuitive inferences that we examined in *Terror Subterra*.

Abductive reasoning, conceived by philosopher Charles Peirce in 1865, concerns reasoning about the *best explanatory hypotheses* for *partial observations*. Examples that compare deduction, induction, and abduction are shown in Table 1. What is remarkable about abductive reasoning is that it is a form of creative reasoning: it *generates new information* that goes beyond what is provided by the premise. Thus, abductive reasoning builds on our imaginative thinking, which, in turn, builds on our rich background knowledge about how the world works. In contrast, the conclusions of deduction and induction do not generate any new information beyond what is already provided in the premise, as these conclusions are only different ways of regurgitating the same or part of the information that is contained in the premise. Generating new hypotheses that explain our partial observations about the world, a cognitive process at the heart of human learning and reasoning, is therefore beyond the conventional scope of formal logic that focuses on truthful conclusions. Although most of our day-to-day reasoning is a form of abductive reasoning, it is relatively less known to most people. For example, Conan Doyle, the author of the Sherlock Holmes canon, mistakenly wrote that Sherlock used deductive reasoning to solve his cases. On the contrary, the key to solving Holmes's mysteries was almost always abductive reasoning, which requires a nontrivial dose of imagination and causal reasoning to generate explanatory hypotheses that may not seem obvious to others. In fact, abductive reasoning is the key to scientific advances as well, since scientific inquiries also require generating new explanatory hypotheses beyond what is already known to the field as truth.

Despite the significance of abduction in human reasoning, relatively few researchers have developed computational systems of abductive reasoning, especially in relation to language-based reasoning. Within the AI logic research communities, language has been very rarely or only minimally integrated into reasoning, as prior research aimed to operate on top of logic-based formalisms detached from natural language. In contrast, within natural language processing (NLP) research communities, a subfield of AI that focuses on human language technologies, questions about intuitive reasoning, commonsense reasoning, and abductive reasoning have by and large been considered to be outside the scope of the field.

Counterfactual reasoning is closely related to abductive reasoning in that they are both cases of nonmonotonic reasoning: that is, logical conclusions are not monotonically true and can be defeasible.[9] Similar to abductive reasoning, counterfactual reasoning has been relatively less studied, and what prior research on counterfactual reasoning there is has been mostly detached from natural language.

Table 1
Examples of Deduction, Induction, and Abduction

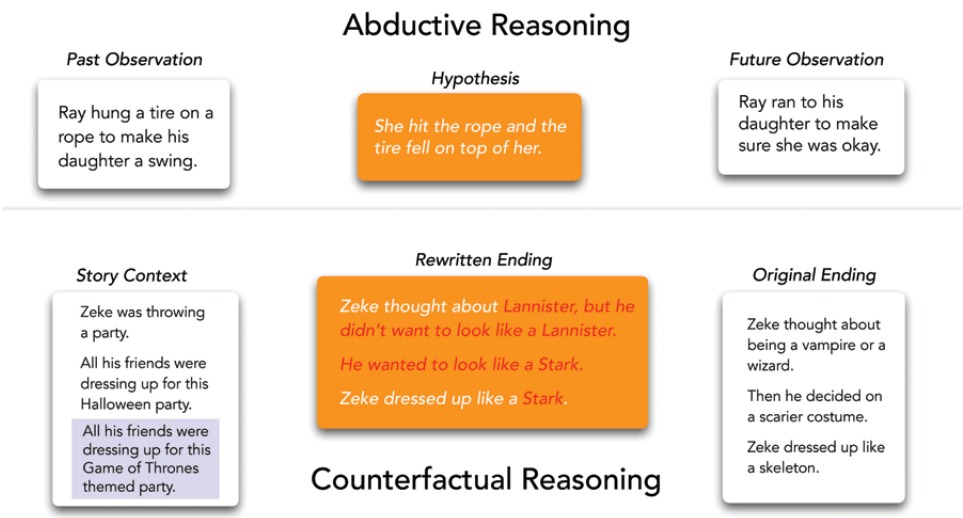
Deduction	Induction	Abduction
There are two monsters running down the tunnel. Jack is the monster in the front.	There is one monster in the tunnel that is running. Another monster enters the tunnel and starts running.	There are two monsters running down the tunnel in sequence.
➔ Jack is running down the tunnel.	➔ All monsters in the tunnel are running.	➔ The one behind is chasing after the one in the front. ➔ The chaser has hostile intentions.

NLP researchers only recently began investigating language-based commonsense reasoning,[10] defeasible inferences,[11] and abductive reasoning,[12] and most recent successes have built on neural language models operating directly with natural language, without formal logical forms.

We have identified the need for designing on-the-fly generative reasoning models through language. But using off-the-shelf language models is not straightforward because generative language models are typically trained for generating language monotonically, such as from left to right for English text. In contrast, abductive and counterfactual reasoning, core abilities of everyday human cognition, require flexible causal reasoning over events that might not be monotonic in time. For example, we might need to condition on the future and reason about the past. Or we might need to condition on both the past and the future to reason about what might have happened in between.

My colleagues and I have recently proposed DeLorean (named after the time-travel machine from *Back to the Future*), a new inference algorithm that can flexibly incorporate both the past and future contexts using only off-the-shelf, left-to-right language models, and no supervision.[13] The key intuition of our algorithm is incorporating the future through "back-propagation," in which we only update the internal representation of the output while fixing the model parameters. By alternating between forward and backward propagation of information, DeLorean can decode the output representation that reflects both the past and future contexts.

Figure 2
Example of DeLorean Reasoning for Abductive *(top)* and Counterfactual Reasoning *(bottom)*

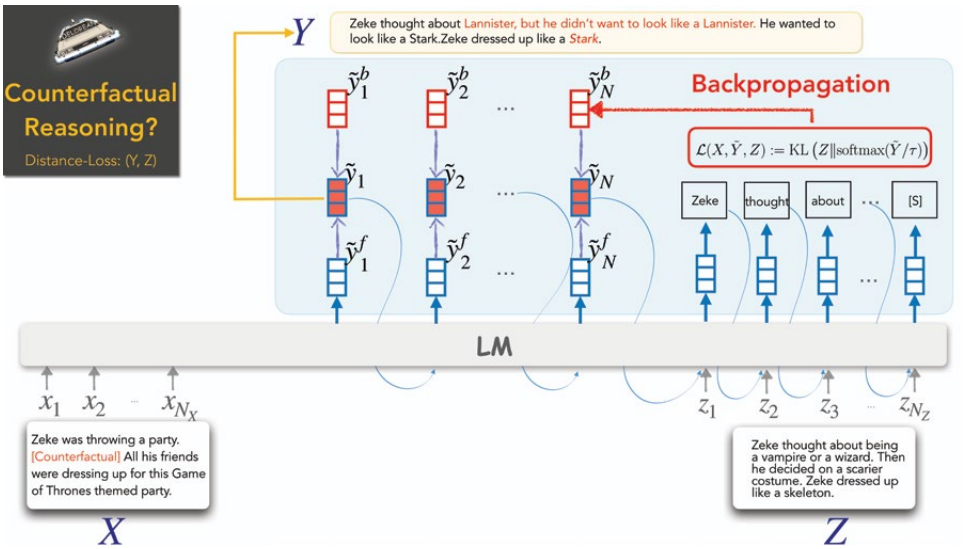


Given the inputs (text boxes on the left and right), DeLorean generates an output (text boxes in the middle). Source: Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, et al., "Abductive Commonsense Reasoning," paper presented at the International Conference on Learning Representations, March 29, 2020; Lianhui Qin, Antoine Bosselut, Ari Holtzman, et al., "Counterfactual Story Reasoning and Generation," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing* (Stroudsburg, Pa.: Association for Computational Linguistics, 2019); and Lianhui Qin, Vered Shwartz, Peter West, et al., "Back to the Future: Unsupervised Backprop-Based Decoding for Counterfactual and Abductive Commonsense Reasoning (DeLorean)," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing* (Stroudsburg, Pa.: Association for Computational Linguistics, 2020).

We have demonstrated that our approach is general and applicable to two nonmonotonic reasoning tasks – abductive text generation and counterfactual story revision – and that DeLorean outperforms a range of unsupervised and some supervised methods based on automatic and human evaluation. Figure 2 illustrates example model outputs, and Figure 3 provides a visual sketch of our method.

COMET, a recent Allen Institute for AI and University of Washington advancement toward commonsense modeling, is another empirical demonstration of on-the-fly generative reasoning through language.[14] COMET is trained using "a large-scale common sense repository of textual descriptions that

Figure 3
Sketch of DeLorean Operations



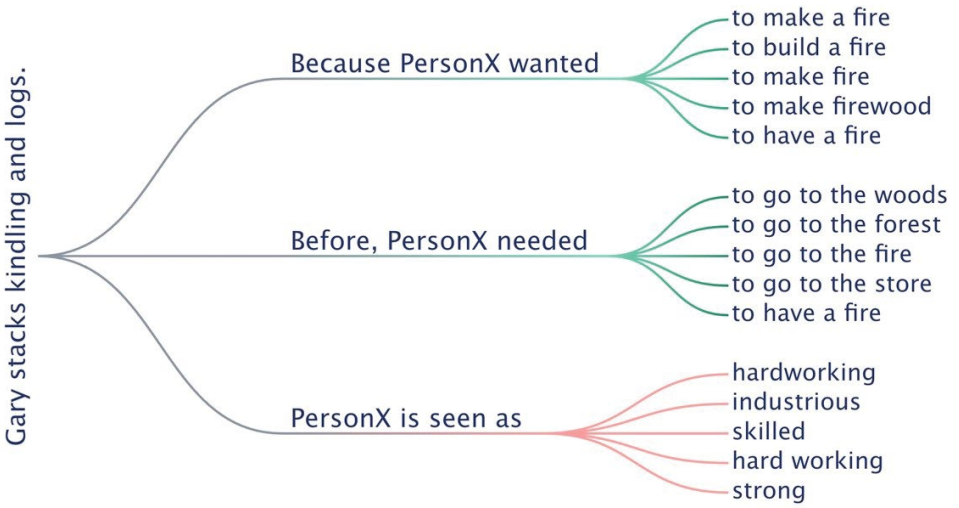
The inputs at the bottom (text boxes X and Z) correspond to the past and the future context on which the DeLorean conditions. The output from DeLorean reasoning is shown at the top of the figure (text box Y).

encode both the social and the physical aspects of common human everyday experiences." But the best way to understand COMET is to experience it for yourself through examples and a live demonstration at https://comet.allenai.org. There you can supply COMET with a statement, and it will predict the subject's relationship with past, future, and present events, characters, and conditions.

Figure 4 shows a COMET prediction given the input "Gary stacks kindling and logs and drops some matches." The model correctly predicts that Gary (that is, PersonX) might want "to start a fire," and before doing so, Gary probably needed "to get a lighter." This particular example was in response to cognitive scientist Gary Marcus's critique on the limitations of neural language models in their commonsense capabilities.[15] Indeed, off-the-shelf neural language models fall far short of robust commonsense intelligence, which motivates the development of commonsense models like COMET.

The key conceptual framework underlying COMET, compared with most commonsense systems from previous decades, is the combination of language-based formalism of commonsense knowledge (as opposed to logic-based formalism)

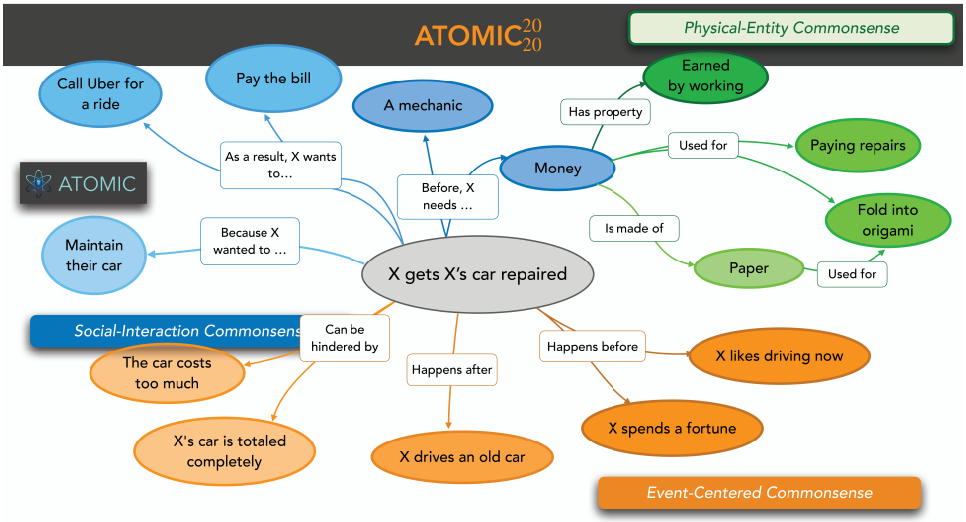
Figure 4
Commonsense Inferences by COMET Given the Input "Gary Stacks Kindling and Logs"



Source: "Commonsense Inferences about People and Events (COMmonsensE Transformers on Atomic2020)," Mosaic Knowledge Graphs, Allen Institute for AI, https://comet.allenai.org.

and on-the-fly generative reasoning over the infinite space of intuitive inferences (as opposed to discriminative prediction over the fixed set of categories). COMET is built on top of ATOMIC, a symbolic knowledge graph that can be viewed as a textbook customized for neural language models to learn commonsense knowledge about how the world works.[16] Analogous to textbooks written for humans, which provide declarative knowledge about a particular topic, ATOMIC is a collection of declarative knowledge focusing on commonsense rules and facts about everyday objects and events. Examples of knowledge encoded in ATOMIC are shown in Figure 5. At the time of writing, ATOMIC draws on more than 1.3 million pieces of commonsense rules and facts. This may sound like a lot, but in reality, 1.3 million pieces of rules and facts are still too limiting to encompass all the trivial commonsense knowledge that we humans hold about the world. Consider that the example of someone stacking kindling and logs is not covered by ATOMIC, nor are Roger Shepard's monsters in a tunnel. Yet COMET, which is trained on ATOMIC, can generalize far beyond the limited scope of symbolic knowledge spelled out in ATOMIC, and can make remarkably accurate commonsense inferences on previously unseen situations, as shown in Figure 4.

Figure 5
Examples of Knowledge Encoded in ATOMIC, the Symbolic Commonsense Knowledge Graph



Source: Jena Hwang, Chandra Bhagavatula, Ronan Le Bras, et al., "(Comet-)Atomic-2020: On Symbolic and Neural Commonsense Knowledge Graphs," paper presented at The Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI-21), February 2–9, 2021.

This generalization power of COMET is achieved through computational melding between neural representation of language and the symbolic representation of commonsense knowledge. Indeed, the empirical success of COMET can be attributed to the blend of neural and symbolic representation of knowledge and the use of language as the representation medium for symbolic knowledge. It is also important to recognize the continuum between knowledge and reasoning. This may seem counterintuitive, as knowledge and reasoning are commonly considered distinct intellectual phenomena. But our computational exploration of language, knowledge, and intuitive reasoning has revealed that, when encountered with a wide spectrum of real-life examples, the boundary between knowledge and reasoning is not clear. More concretely, when we reason about the intent of "Gary stacking kindling and logs," our reasoning relies on our memorized commonsense knowledge about what people typically do with kindling and logs. Conversely, frequent patterns of commonsense reasoning about the intents and mental states of people, the causes and effects of events, and the preconditions and postconditions

of events all become integral parts of our memorized knowledge about how the world works. In sum, COMET demonstrates a neuro-symbolic blend between language, knowledge, and reasoning as a new path toward commonsense AI. Without this mix, the remarkable generalization power of COMET to flexibly reason about previously unseen situations would have been unattainable.

While the curious case of commonsense intelligence remains far from solved, the investigation thus far has made considerable progress toward insights that may crack the old mystery. Like in any good mystery, there are many surprises still to come, but recent projects have meaningfully built on the key ideas behind ATOMIC and COMET to blend language, knowledge, and reasoning; I will introduce two.

A new algorithmic framework called Symbolic Knowledge Distillation has enabled us to distill symbolic knowledge from neural networks (GPT-3 in particular) algorithmically.[17] In a nutshell, instead of humans writing the symbolic commonsense knowledge graph, such as ATOMIC, to teach machines with, machines can now author their own knowledge graph with which to teach themselves. Moreover, the resulting machine-authored ATOMIC can exceed, for the first time, the human-authored counterpart in all criteria: scale, quality, and diversity. This development foreshadows a great many adventures ahead of us.

But what would it take to teach a machine to behave ethically? Delphi, the second project, is a prototype commonsense morality and norms model. While some broad ethical rules are captured by straightforward statements ("thou shalt not kill"), applying such rules to real-world situations is far more complex. For example, while "helping a friend" is generally a good thing to do, "helping a friend spread fake news" is not.

Delphi is designed to reason about simple ethical situations (you can submit your own for judgment at https://delphi.allenai.org/).[18] As shown in Figure 6, making an ethical judgment of a given situation requires understanding a broad range of ethical and social norms, and complex reasoning to calibrate across competing values (such as killing a bear versus pleasing your child).

Delphi demonstrates the promises of language-based commonsense moral reasoning, with up to 80–92 percent accuracy, as vetted by humans. This is in stark contrast to the off-the-shelf performance of GPT-3 of 52.3 percent accuracy, which suggests that massive scale alone does not endow pretrained neural language models with human values.

Thus, Delphi is taught with the Commonsense Norm Bank, a moral textbook customized for machines that compiles 1.7 million examples of people's ethical judgments on diverse everyday situations. The Commonsense Norm Bank is analogous to ATOMIC in that both are symbolic knowledge bases/textbooks used to teach machines. The scope of the Norm Bank overlaps with but goes much further than that of ATOMIC: the former focuses on social and ethical norms for everyday

Figure 6
Delphi Judgments on Previously Unseen Questions



Source: Delphi, "Ask Delphi," Allen Institute for AI, https://delphi.allenai.org/ (accessed December 8, 2021).

situations, including problems on equity, in order to teach AI against racism or sexism.

While Delphi shows promise, the Delphi study has also revealed major limitations of neural models for their unfiltered bias and harms. The study also opens up new research questions, including how we can revise the Commonsense Norm Bank so its examples represent more diverse cultural norms.[19]

Delphi is an emblematic project toward the bigger goal of teaching AI to behave in more inclusive, ethically informed, and socially aware manners when interacting with humans. As AI systems become increasingly integral in people's everyday lives, it becomes a priority that they learn to respect human values and behave ethically. However, AI systems are not, and should never be, used as moral authorities or sources of advice on human ethics. The fact that AI learns to interact with humans ethically does not make the AI a moral authority over humans, just like a human who tries to behave ethically does not become the moral authority over other people.

We have discussed the importance of deep integration of language toward commonsense AI, as well as why numerous past attempts based on logic-based formalisms, despite their intellectual appeal, did not empirically model the rich scope of intuitive reasoning that humans find trivial for everyday objects and events. While the research highlighted in this essay demonstrates potential new paths forward, we are far from solving commonsense AI. Numerous open research questions remain, including computational mechanisms to ensure consistency and interpretability of commonsense knowledge and reasoning, deep representational integration between language and perception for multimodal reasoning, new learning paradigms for abstraction and analogies, and advanced learning methods for interactive and lifelong learning of knowledge and reasoning.

ABOUT THE AUTHOR

Yejin Choi is the Brett Helsel Professor at the Paul G. Allen School of Computer Science and Engineering at the University of Washington and Senior Research Manager at the Allen Institute for Artificial Intelligence, where she oversees the project Mosaic. She has recently published in the proceedings of such conferences as Advances in Neural Information Processing Systems, the Association for Computational Linguistics, and the AAAI Conference on Artificial Intelligence.

ENDNOTES

[1] Gary Marcus and Ernest Davis, *Rebooting AI: Building Artificial Intelligence We Can Trust* (New York: Vintage, 2019); and Ernest Davis and Gary Marcus, "Commonsense Reasoning and Commonsense Knowledge in Artificial Intelligence," *Communications of the ACM* 58 (9) (2015): 92–103.

[2] MOSAIC, The Allen Institute for Artificial Intelligence, https://mosaic.allenai.org/.

[3] Maarten Sap, "ACL 2020 Commonsense Tutorial (T6)," https://homes.cs.washington.edu/~msap/acl2020-commonsense/.

[4] Matt Turek, "Machine Common Sense (MCS)," Defense Advanced Research Projects Agency, https://www.darpa.mil/program/machine-common-sense.

[5] Roger N. Shepard, *Mind Sights: Original Visual Illusions, Ambiguities, and Other Anomalies, with a Commentary on the Play of Mind in Perception and Art* (New York: W. H. Freeman and Co., 1990); and Hugo Mercier and Dan Sperber, *The Enigma of Reason* (Cambridge, Mass.: Harvard University Press, 2017).

[6] Douglas Hofstadter and Emmanuel Sander, *Surfaces and Essence: Analogy as the Fuel and Fire of Thinking* (New York: Basic Books, 2013).

[7] Tom Brown, Benjamin Mann, Nick Ryder, et al., "Language Models are Few-Shot Learners," *Advances in Neural Information Processing Systems* 33 (2020).

[8] Mercier and Sperber, *The Enigma of Reason*.

[9] "Non-monotonic Logic," Stanford Encyclopedia of Philosophy, substantive revision April 20, 2019, https://plato.stanford.edu/entries/logic-nonmonotonic/.

[10] Lianhui Qin, Antoine Bosselut, Ari Holtzman, et al., "Counterfactual Story Reasoning and Generation," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing* (Stroudsburg, Pa.: Association for Computational Linguistics, 2019); Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi, "SWAG: A Large-Scale Adversarial Dataset for Grounded Commonsense Inference," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (Stroudsburg, Pa.: Association for Computational Linguistics, 2018); and Rowan Zellers, Ari Holtzman, Yonatan Bisk, et al., "HellaSwag: Can a Machine Really Finish Your Sentence?" in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (Stroudsburg, Pa.: Association for Computational Linguistics, 2019).

[11] Rachel Rudinger, Vered Shwartz, Jena D. Hwang, et al., "Thinking Like a Skeptic: Defeasible Inference in Natural Language," in *Findings of the Association for Computational Linguistics: EMNLP 2020* (Stroudsburg, Pa.: Association for Computational Linguistics, 2020).

[12] Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, et al., "Abductive Commonsense Reasoning," paper presented at the International Conference on Learning Representations, March 29, 2020.

[13] Lianhui Qin, Vered Shwartz, Peter West, et al., "Back to the Future: Unsupervised Backprop-Based Decoding for Counterfactual and Abductive Commonsense Reasoning (DeLorean)," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing* (Stroudsburg, Pa.: Association for Computational Linguistics, 2020).

[14] Antoine Bosselut, Hannah Rashkin, Maarten Sap, et al., "COMET: Commonsense Transformers for Knowledge Graph Construction," in *Proceedings of the 57th Annual Meeting of*

the Association for Computational Linguistics (Stroudsburg, Pa.: Association for Computational Linguistics, 2019); and Jena Hwang, Chandra Bhagavatula, Ronan Le Bras, et al., "(Comet-)Atomic-2020: On Symbolic and Neural Commonsense Knowledge Graphs," paper presented at The Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI-21), February 2–9, 2021.

[15] Gary Marcus, @GaryMarcus, Twitter, October 26, 2019, 11:55 p.m., https://twitter.com/GaryMarcus/status/1188303158176403457; Gary Marcus, @GaryMarcus, Twitter, October 26, 2019, 11:57 p.m., https://twitter.com/YejinChoinka/status/1188312418562134016; Marcus and Davis, *Rebooting AI*; and Ernest Davis and Gary Marcus, "Commonsense Reasoning and Commonsense Knowledge in Artificial Intelligence," *Communications of the ACM* 58 (9) (2015): 92–103.

[16] Hwang et al., "(Comet-)Atomic-2020"; and Maarten Sap, Ronan LeBras, Emily Allaway, et al., "ATOMIC: An Atlas of Machine Commonsense for If-Then Reasoning," in *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence* (Palo Alto, Calif.: AAAI Press, 2019).

[17] Peter West, Chandra Bhagavatula, Jack Hessel, et al., "Symbolic Knowledge Distillation: From General Language Models to Commonsense Models," arXiv (2021), https://arxiv.org/abs/2110.07178; and Yannic Kilcher, "Symbolic Knowledge Distillation: From General Language Models to Commonsense Models (Explained)," YouTube video, uploaded October 24, 2021, https://www.youtube.com/watch?v=kP-dXK9JEhY.

[18] Liwei Jiang, Jena D. Hwang, Chandra Bhagavatula, et al., "Delphi: Towards Machine Ethics and Norms," arXiv (2021), https://arxiv.org/abs/2110.07574; and Liwei Jiang, Jena D. Hwang, Chandra Bhagavatula, et al., "Towards Machine Ethics and Norms," AI2 blog, November 3, 2021, https://medium.com/ai2-blog/towards-machine-ethics-and-norms-d64f2bdde6a3.

[19] Ibid.

Language & Coding Creativity

Ermira Murati

Machines are gaining understanding of language at a very rapid pace. This achievement has given rise to a host of creative and business applications using natural language processing (NLP) engines, such as OpenAI's GPT-3. NLP applications do not simply change commerce and literature. They raise new questions about how human beings relate to machines and how that symbiosis of communication will evolve as the future rushes toward us.

Every writer has a unique aesthetic in the way they order words. The nuances of applied language, or voice, mark one of the countless fingerprints of human creativity. Decoding the secrets of this language sits at the frontier of artificial intelligence: how to build machines that truly understand not only language at a human level, but produce human-grade responses too.

Take the following excerpt of a poem: "For you are the most beautiful thing we have in this world / I love your graceful symmetry, your simplicity and clarity / You are the song of the Universe, a cosmic lullaby / You are the poetry of nature, written with light and electricity / You are the music of the spheres, played on a harp made of vacuum."[1] The directness, the imagery, the fearless affection, one might believe the words to be Pablo Neruda's. But Neruda is only part of the answer. An artificial intelligence system known as GPT-3 (Generative Pre-trained Transformer 3), built by the research laboratory OpenAI, scanned an enormous corpus of language data, including Neruda's verses, and built probabilistic relationships of tremendous fidelity between his use of nouns, verbs, adjectives, objects, and all the mechanics of a poem. Consequently, GPT-3 could independently generate this brand-new poem in its own voice.

For decades, some visionary scientists have predicted this level of intricacy from a machine. The user only had to give GPT-3 a prompt, or rather, inspiration: "The following is a poem about Maxwell's equations in the style of poet Pablo Neruda." From that instruction, the machine could pull from its brain of data to not only grasp aspects of Maxwell's foundational electromagnetic equations but present them in Neruda's style.

This approach to AI is known as large language models and its applications are spreading across the arts, sciences, and business. Those overseeing the code and training of the machine are becoming authors – and editors – of a new collective

© 2022 by Ermira Murati
Published under a Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) license
https://doi.org/10.1162/DAED_a_01907

language. Train a machine with a corpus of text and it can answer customer service questions, describe the plays of a football game, compose an essay, or write a program based on a description of its function. The applications are not only becoming more integral to commerce and our daily lives but are spawning questions about the nature of language. Why do certain aesthetics ring true while other deployments of language feel empty or fake, even when the grammar is perfect? We can understand more about our own processes of thought by understanding how a machine decides to use language.

Technology, culture, civilization: none comes into being without language. Language is both a high point and the foundation of human intelligence. Yet there is a bind: What are languages exactly? How do they work? We might think of language as a reaction to context and surroundings. But if we cannot write out the rules of language, how do we teach it to a machine? This problem has captivated thinkers for a century, and the answers are now starting to appear.

What is a thought? And how is experiencing a thought different from experiencing a memory or an idea? It is difficult to understand; to borrow from philosophy, digging into the roots of consciousness or any working of the mind starts to feel like trying to see our own eyes or bite our own teeth. Staring into space or perspiring over a pad of paper, thoughts seem to work less like a hard disk and more like a wind, arriving and departing without an obvious explanation.

Our thoughts manifest through action and emotion but are communicated through language. Charles Darwin put language on the razor's edge between an instinct and a skill. A human baby starts babbling almost instantly – call it innately – yet takes years to engage in higher level conversations around them. At the same time, all languages are learned, whether directly or passively. That learning takes years of repetition. Whereas a toddler can hold a casual conversation, they need another decade before writing structured paragraphs.

Darwin saw the drive to acquire language as "the instinctive tendency to acquire an art," to communicate by some medium.[2] No baby has ever needed a book of grammar to learn a language. They absorb what they hear and through the maze of the mind play it back. People spend their lives speaking exquisitely without understanding a subjunctive clause or taking a position on split infinitives. A child learns by experiencing patterns, learning what is most likely to make sense in a new context. To paraphrase Ralph Waldo Emerson, this information shapes who we become, much like every meal we have eaten.

The mechanics of the mind are still a mystery. The nuances of a writer's voice and creativity in general are no exception. Neuroscientists can observe now that certain neurons light up when certain actions occur, be it throwing a ball or giving directions. The order in which these neurons light up and the connections between

them seem to dictate – or "code" – what the owner of the brain experiences. Researchers have discovered neurons that fire in response to extremely, perhaps oddly, specific subjects, such as Halle Berry. The same neurons will activate when reading the actress's name or seeing a picture of her, yet remain silent for anyone else.[3]

We now design machines to mimic a similar logic. If you are shown a photo of Halle Berry and asked who are you looking at? you fill in the blank with the answer "most likely" to be correct. Every word you output is the most likely word to be the "right" answer, or to ring true. In a system like GPT-3, these blanks are known as parameters. As the machine goes along creating text, it asks itself over and over (and over): with the text I have been given, what word is most likely to come next? GPT-3 reaches answers based on the patterns it identifies from the existing usage of human language creating a map, the same way we might piece characters together when writing essays or code. With this mathematical representation of patterns, GPT-3 can carry out many tasks, such as generating computer code, translating language, classifying information, searching semantically through a large corpus of text, even though it was built to do just one thing: predict the next word in a sequence of words.

For a long time computers were thought best suited for "simple repetitive tasks." There is little creativity in arithmetic, so machines make excellent calculators, accountants, and modelers. When it came to training a machine in language, there was a critical bind: Whose language? And if you agree on a model of language, how do you deliver that information?

The answer to date has been to cast a wide net in those languages and exercise repetition and volume to deliver the information. GPT-3 ingested about one trillion words from the Internet, which consist of instances of text from countless sources.[4] It built relationships between these parameters, with GPT-3 developing a kind of reactive memory in the process. Now a collaborator can give GPT-3 a task and the system understands it, such as "make a Pablo Neruda poem as an ode to Planck's equations":

In the darkness
The uncertainty
Of the void
There is a light
The only one
That shines.
It is a tiny light
Yet powerful
It is the photon
Which carries
The secret
And science
Of the stars
It is the answer
The light of the dark
The dark of the light
It is
The equation
Of everything.

The origins of predicting what word comes next has roots in Russian literature. Scan the letters of this text and you will notice that three consonants rarely appear in a row. Four consonants, hardly ever. Russian mathematician Andrey Markov began quantifying this phenomenon in 1913 by writing out the first twenty thousand letters of Alexander Pushkin's novel *Eugene Onegin*. Converting what we intuitively know into numbers, Markov showed how the preceding letters dictate the probability of what comes next. But Markov could only compare the rates of vowels and consonants. In that day, it would have been impossible to map on graph paper all letters and their respective frequencies in relation to the rest of the text in two and three letter combinations. Today, machines answer these questions in an instant, which is why we see so many applications interfacing with conversational language. Rather than predicting the next letter, GPT-3 predicts what word comes next by reviewing the text that came before it.

Human speech works this same way. When you walk into a room and say "I need a ____," a relatively narrow list of words would make sense in the blank. As the context becomes more detailed – for instance, walking into a kitchen covered in mud – that list shrinks further. Our minds develop this sorting naturally through experiences, but to train GPT-3's mind, the system has to review hundreds of billions of different data points and work out the patterns among them.

Since Markov's contributions, mathematicians and computer scientists have been laying the theoretical groundwork for today's NLP models. But it took recent advances in computing to make these theories reality: now processors can handle billions of inputs and outputs in milliseconds. For the first time, machines can perform any general language task. From a computer architecture sense, this has helped unify NLP architectures. Previously, there were myriad architectures across mathematical frameworks – recurrent neural networks, convolutional neural networks, and recursive neural networks – built for specific tasks. For a machine answering a phone call, previously, the software relied upon one mathematical framework to translate the language, another to dictate a response. Now, GPT architecture has unified NLP research under one system.

GPT-3 is the latest iteration of generative pretrained transformer models, which were developed by scientists at OpenAI in 2018. On the surface, it may be difficult to see the difference between these models and more narrow or specific AI models. Historically, most AI models were trained through supervised machine learning, which means humans labeled data sets to teach the algorithm to understand patterns. Each of these models would be developed for a specific task, such as translating or suggesting grammar. Every model could only be used for that specific task and could not be repurposed even for seemingly similar applications. As a result, there would be as many models as there were tasks.

Transformer machine learning models change this paradigm of specific models for specific tasks to a general model that can adapt to a wide array of tasks. In

2017, researchers Alec Radford, Rafal Jozefowicz, and Ilya Sutskever identified this opportunity while studying next character prediction, in the context of Amazon reviews, using an older neural network architecture called the LSTM. It became clear that good next character prediction leads to the neural network discovering the sentiment neuron, without having been explicitly told to do so. This finding hinted that a neural network with good enough next character or word prediction capabilities should have developed an understanding of language.

Shortly thereafter, transformers were introduced. OpenAI researchers immediately saw their potential as a powerful neural network architecture, and specifically saw the opportunity to use it to study the properties of very good next word prediction. This led to the creation of the first GPT: the transformer language model that was pretrained on a large corpus of text, which achieved excellent performance on every task using only a little bit of finetuning. As OpenAI continued to scale the GPT, its performance, both in next word prediction and in all other language tasks, kept increasing monotonically, leading to GPT-3, a general purpose language engine.

In the scope of current AI applications, this may at first seem a negligible difference: very powerful narrow AI models can complete specific tasks, while a GPT architecture, using one model, can also perform these separate tasks, to similar or better results. However, in the pursuit of developing true, human-like intelligence, a core tenet is the ability to combine and instantly switch between many different tasks and apply knowledge and skills across different domains. Unified architectures like GPT will therefore be key in advancing AI research by combining skills and knowledge across domains, rather than focusing on independent narrow tasks.

Humans also learn language through other senses: watching, smelling, touching. From the perspective of a machine, these are different modes of training. Today, we try to simulate this human way of learning by not only training a machine's cognitive processing on words, but on images and audio too. We use this multimodal approach to teach a machine how words relate to objects and the environment. A taxi is not just the letters T-A-X-I, but a series of sounds, a pixel pattern in digital photos, a component of concepts like transportation and commerce. Weaving these other modes into a machine broadens the applications developers can build, as the machine's brain is able to apply its knowledge across those different modes as well.

An example is designing a web page. Every business struggles with keeping its site up-to-date, not only updating text, photos, and site architectures, but also understanding how to code the CSS and HTML. This is both time-consuming and costly. Developers have demonstrated that GPT-3 can understand layout instructions and build the appropriate mockups, for instance, when you tell it to "add a header image with an oak tree and my contact information below." Under the hood, GPT-3 is transforming between the vast arrays of text and the vast array of objects. The result is that a person without any website-building experience can have a piece of working HTML in seconds.

Figure 1
Codex-Generated Completion

```
You will be given a string of words separated by
    commas or spaces. Your task is
to split the string into words and return an array
    of the words.

For example:
words_string("Hi, my name is John") == ["Hi", "my",
     "name", "is", "John"]
words_string("One, two, three, four, five, six") ==
     ["One", "two", "three", "four", "five", "six"]
"""

### COMPLETION 1 (CORRECT): ###
words = []
splitted_word = ""
for char in s:
    if char == "␣":
        if splitted_word != "":
            words.append(splitted_word)
            splitted_word = ""
    elif char != ",":
        splitted_word += char
if splitted_word != "":
    words.append(splitted_word)
return words
```

Drawing on publicly available code on GitHub, the GPT model is able to come up with the correct solution. Source: Mark Chen, Jerry Tworek, Heewoo Jun, et al., "Evaluating Large Language Models Trained on Code," arXiv (2021), https://arxiv.org/abs/2107.03374.

The next stage is using a GPT model in more advanced programming languages such as Python. Programmers are often thrust into coding projects in which they do not know the logic of everything that has been written already, like having to continue writing a half-finished novel. Usually, programmers spend substantial amounts of time and effort getting up to speed, whereas Codex (Figure 1), a GPT language model fine-tuned on publicly available code from the development platform GitHub, can scan millions of lines of code and describe to the programmer the function of each section.[5] This saves countless hours of work, but also allows these specialized professionals to focus on creativity and innovation rather than menial tasks.

The next step would be the "writing" of physical objects. For instance, industrial designers are constantly creating and testing new forms and functionalities of products. Imagine they want to build a chair in the shape of an avocado, which requires having both an understanding of the functionality of a chair and the form of an avocado. OpenAI used a 12-billion parameter version of GPT-3 known as DALL·E and trained it to generate images from text descriptions, using a data set of text-image pairs. As a result, DALL·E gained a certain understanding of the relationship between text and images. When DALL·E was then prompted to suggest designs for "an armchair in the shape of an avocado" it used its understanding to propose designs (Figure 2).[6]

Figure 2
DALL·E Successor Iterates on the Text Prompt: "An Armchair in the Shape of an Avocado"



Figure 2, continued
DALL·E Successor Iterates on on the Text Prompt: "A Fox in the Style of *Starry Night* by Van Gogh"



Asked to create designs of an avocado chair and Van Gogh-inspired fox, this model drew on its understanding of the functions of chairs and Van Gogh's painting style in *Starry Night*. None of these designs or images existed until the model created them. Source: Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, et al., "Zero-Shot Text-to-Image Generation," arXiv (2021), https://arxiv.org/abs/2102.12092; and Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, et al., "GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models," arXiv (2021), https://arxiv.org/abs/2112.10741.

DALL·E was able to create images that are instantly recognizable as avocado chairs, even though we might struggle ourselves to create instantly such a design. The model is able not only to generate original creative output, as avocado chairs are not a common product easily found and copied elsewhere, but also adheres in its designs to the implicit constraints of form and functionality associated with avocados and chairs.

This does not put human designers out of a job. Rather, they gain a team of assistants to take on their most rote tasks, allowing them instead to focus on curating and improving on good ideas or developing their own. In the same way GPT-3 summarizing, explaining, and generating Python code opens up programming to nonprogrammers, such iterative design opens up avenues for nondesigners. A small business or individual designer now has access to capabilities that otherwise may have only been accessible to large organizations.

There are a multitude of applications in which transformer models can be useful, given that they can not only understand but also generate output across these different modes. GPT-3 has already been used for understanding legal texts through semantic search tools, helping writers develop better movie scripts, writing teaching materials and grading tests, and classifying the carbon footprint of purchases.

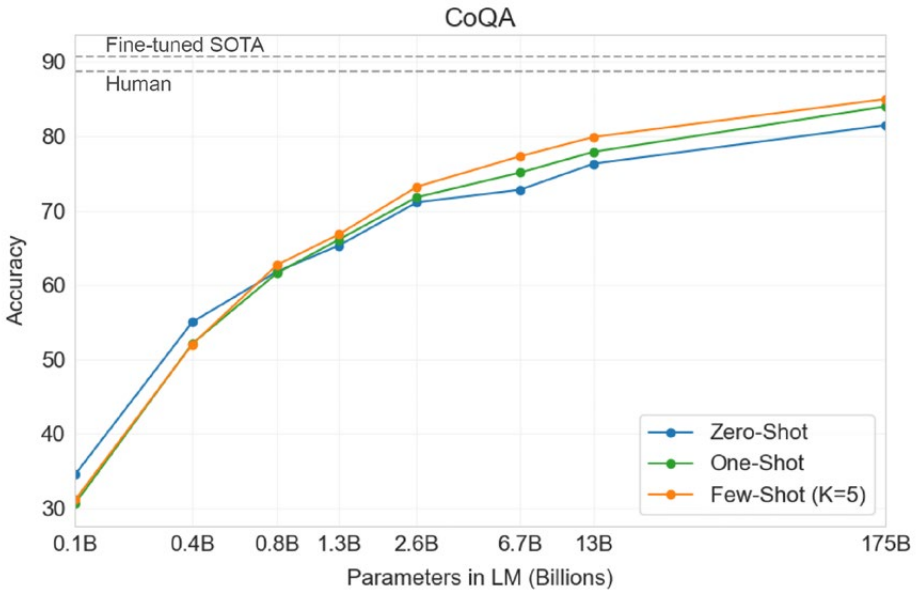
Tracking the progress of GPT models over the past few years, we can see what the future might bring in terms of model performance. GPT-2 was a one-and-a-half-billion-parameter model trained on forty gigabytes of data, which is an amount of text about eight thousand times larger than the collected works of Shakespeare. GPT-3, more than one hundred times bigger, comes close to human comprehension on complex reading tests (see Figure 3). As we move forward in both model complexity and the size of the data sets, we believe these models will move ever closer to human benchmarks.

At the same time, as they are tested and applied more extensively, we find limitations in these models. For instance, GPT-3 shows notable weakness in generating long passages, struggling with self-repetition, non sequiturs, and incoherence. It also struggles with seemingly commonsense questions, such as: "If I put cheese in the fridge, will it melt?"

There is always a duality to powerful technological disruptions. The advent of network computing in 1989 paved the way for the Internet. Tim Berners-Lee envisioned the Internet as "a collaborative space where you can communicate through sharing information."[7] With freedom of access to all knowledge and boundaries dissolved, the Internet opened Pandora's box. Next to the many positives, it also provides thoroughfares for misinformation, trolling, doxing, crime, threats, and traumatizing content.

It would be naive to consider GPT-3's optimal impact without reflecting on what pitfalls might lie before us. GPT-3 is built to be dynamic and require little data

Figure 3
GPT-3 Results on CoQA Reading Comprehension Task



GPT-3 175B is only a few points behind the accuracy of human performance and state-of-the-art fine-tuned models. Source: Tom B. Brown, Benjamin Mann, Nick Ryder, et al., "Language Models are Few-Shot Learners," arXiv (2020), https://arxiv.org/abs/2005.14165.

to perform a task, but the system's experience will color its future work. This experience will always have holes and missing pieces. Like human beings, machines take inputs and generate outputs. And like humans, the output of a machine reflects its data sets and training, just as a student's output reflects the lessons of their textbook and teacher. Without guidance, the system will start to show blind spots, the same way a mind focused on a single task can become rigid compared with a mind performing many tasks and gathering a wide variety of information.

In AI, this phenomenon is broadly known as bias, and it has consequences. For instance, a health care provider may use an NLP model to gather information on new patients and may train this model on the responses from a certain demographic distribution. A new patient outside that distribution might be poorly assisted by this system, causing a negative experience for someone needing help.

More generally, powerful language models can increase the efficacy of socially harmful activities that rely on text generation. Examples include misinformation, abuse of legal and governmental processes, spam, and phishing. Many of these

harmful activities are limited by having enough human talent and bandwidth to write texts and distribute them, whereas with GPT models, this barrier is lowered significantly.

Moreover, generative language models suffer from an issue shared by many humans: the inability to admit a lack of knowledge or expertise. In practical terms, language models always generate an answer – even if it is nonsensical – instead of recognizing that it does not have sufficient information or training to address the prompt or question.

As NLP models continue to evolve, we will need to navigate many questions related to this duality. Developers are already writing books using machines processing what they experience in the world. How do we draw the boundary between the creator and the code? Is the code a tool or an extension of the mind? These questions go well beyond the arts. How long until machines are writing scientific papers? Machines are already conducting large sections of experiments autonomously. Language can also say a lot about our confidence or mood. Do we want a company basing product recommendations off what we thought was an innocent interaction? How do creators, users, and uses create bias in a technology?

For the first time, we are using artificial intelligence tools to shape our lives. GPT-3 has shown that large language models can possess incredible linguistic competence and also the ability to perform a wide set of tasks that add real value to the economy. I expect these large models will continue to become more competent in the next five years and unlock applications we simply cannot imagine today. My hope is if we can expose models to data similar to those absorbed by humans, they should learn concepts in ways that are similar to human learning. As we make models like GPT-3 more broadly competent, we also need to make them more aligned with human values, meaning that they should be more truthful and harmless. Researchers at OpenAI have now trained language models that are much better at following user intentions than GPT-3, while also making them more honest and harmless. These models, called InstructGPT, are trained with humans in the loop, allowing humans to use reinforcement to guide the behavior of the models in ways we want, amplifying good results and inhibiting undesired behaviors.[8] This is an important milestone toward building powerful AI systems that do what humans want.

It would not be fair to spend all these words discussing GPT-3 without giving it the chance to respond. I asked GPT-3 to provide a parting thought in response to this essay:

> There is a growing tension between the roles of human and machine in creativity and it will be interesting to see how we resolve them. How we learn to navigate the "human" and "machine" within us will be a defining question of our time.

Artificial intelligence is here to stay, and we need to be ready to embrace it.

ABOUT THE AUTHOR

Ermira Murati is the Senior Vice President of Research and Product at OpenAI, advancing the company's mission to ensure that artificial general intelligence benefits all of humanity. Ermira and her teams are pushing the frontiers of what neural networks can do, seeking to better understand the behavior of powerful AI systems, make them safer, and align them with human intentions and human values.

ENDNOTES

[1] Samhan Salahuddin, "A Wild Adventure With GPT-3–Featuring Indian Mythology and Neruda," Pickled Brains, April 2, 2021, https://pickledbrains.substack.com/p/a-wild-adventure-with-gpt-3.

[2] Charles Darwin, *The Descent of Man, and Selection in Relation to Sex* (London: John Murray, 1874).

[3] Gabriel Goh, Nick Cammarata, Chelsea Voss, et al., "Multimodal Neurons in Artificial Neural Networks," OpenAI, March 4, 2021, https://distill.pub/2021/multimodal-neurons/.

[4] Tom B. Brown, Benjamin Mann, Nick Ryder, et al., "Language Models are Few-Shot Learners," arXiv (2020), https://arxiv.org/abs/2005.14165.

[5] Mark Chen, Jerry Tworek, Heewoo Jun, et al., "Evaluating Large Language Models Trained on Code," arXiv (2021), https://arxiv.org/abs/2107.03374.

[6] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, et al., "Zero-Shot Text-to-Image Generation," arXiv (2021), https://arxiv.org/abs/2102.12092.

[7] "Web's Inventor Gets a Knighthood," BBC, updated December 31, 2003, http://news.bbc.co.uk/1/hi/technology/3357073.stm.

[8] Long Ouyang, Jeff Wu, XuJiang, et al., "Training Language Models to Follow Instructions with Human Feedback," OpenAI, 2022, https://cdn.openai.com/papers/Training_language_models_to_follow_instructions_with_human_feedback.pdf.

Non-Human Words: On GPT-3 as a Philosophical Laboratory

Tobias Rees

In this essay, I investigate the effect of OpenAI's GPT-3 on the modern concept of the human (as alone capable of reason and language) and of machines (as devoid of reason and language). I show how GPT-3 and other transformer-based language models give rise to a new, structuralist concept of language, implicit in which is a new understanding of human and machine that unfolds far beyond the reach of the categories we have inherited from the past. I try to make compelling the argument that AI companies like OpenAI, Google, Facebook, or Microsoft effectively are philosophical laboratories (insofar as they disrupt the old concepts/ontologies we live by) and I ask what it would mean to build AI products from the perspective of the philosophical disruptions they provoke: can we liberate AI from the concept of the human we inherited from the past?

In May 2020, OpenAI released GPT-3 (Generative Pre-trained Transformer 3), an artificial intelligence system based on deep learning techniques that can generate text. GPT-3's interface invites a user to provide the AI system with a bit of text and then, based on the prompt, GPT-3 writes. GPT-3 can write short stories, novels, reportages, scientific papers, code, and mathematical formulas. It can write in different styles and imitate the style of the text prompt. It can also answer content-based questions (that is, it learns the content of texts and can articulate this content). It can translate text from almost any language into almost any other; and it can provide summaries of lengthy passages.

The quality of GPT-3's output is remarkable, often impressive. As many critiques have pointed out, GPT-3 makes silly errors that no human would ever make. And yet GPT-3's translations often exceed translations done by humans and capture even subtle differentiations and wordplays; the summaries are almost always concise; and the text it generates on the basis of prompts is most often surprisingly consistent: GPT-3 can mimic the style of an author to such a degree that is nearly impossible to determine whether the text was written by a given author or by GPT-3.

How can we relate to GPT-3? Or to the many other, often equally powerful large language models (LLMs) built over the last few years: Google's BERT, LaMDA, and Wordcraft; Microsoft's Megatron-Turing Natural Language Generation;

© 2022 by by Tobias Rees
Published under a Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) license
https://doi.org/10.1162/DAED_a_01908

Inspur's YUAN 1.0; Huawei's PanGu-Alpha; Naver's HyperCLOVA; or Sberbank's various Russian models, most notably ruRoBERTa-large?

I have come to think of the development of GPT-3 and its kin as a far-reaching, epoch-making philosophical event: the silent, hardly noticed undoing of the up-until-now exclusive link between humans and words.

The consequences of this undoing are sweeping: the entire modern world – the modern experience of what it is to be human, as well as the modern understanding of reality – is grounded in the idea that we humans are the only talking thing in a world of mute things.

No longer.

Philosophical Stakes

At the beginning of the seventeenth century, a remarkable transformation in our understanding of language took place. Up until that time, the comprehensions of the world as described by Plato and Aristotle had largely remained in place. Most humans still experienced themselves, in accordance with the writings of the Greek philosophers, to be living in a God-given nature-cosmos in which everything – including the human thing – had a well-defined role.

Nature – a metaphysical ground – was all there was.

The particular role of humans in this nature-cosmos was defined by their having language. The assumption was that at the beginning of time, humans received a spark of the divine *logos* that gave things their eternal essence or names, of which the visible world was a mere reflection. This divine gift not only enabled humans to communicate with one another, it also gave them access, via contemplation (a practice that consists in applying logos to itself), to the true names of things and thus to the eternal order of the real.

Around 1600, the ancient, medieval nature-cosmos began to break open. Within a few short decades, the comprehension of reality – the structure of experience of what it is to be human – underwent a remarkably far-reaching change. And at the center of this change was language.

If until then language was a divine gift that enabled humans to know the eternal essence/names of things, then now language became the human unique power to name things and to thereby order and know them and bring them under human control. If language had hitherto defined the role of humans *in* the nature-cosmos, then language was now what set them apart from what was increasingly considered to be *mere* nature: nature was no longer understood and experienced as a divine cosmos but as the *other* of the human, as the nonhuman realm of animals and plants, as mere matter organized in mechanical principles.

The exemplary place where this new concept of language – of humans – is articulated is René Descartes's *Discourse on the Method*, published anonymously in 1630.

> For it is a very remarkable thing that there are no humans, not even the insane, so dull and stupid that they cannot put words together in a manner to convey their thoughts. On the contrary, there is no other animal however perfect and fortunately situated it may be, that can do the same. And this is not because they lack the organs, for we see that magpies and parrots can pronounce words as well as we can, and nevertheless cannot speak as we do, that is, in showing that they think what they are saying. On the other hand, even those humans born deaf and dumb, lacking the organs which others make use of in speaking . . . usually invent for themselves some signs by which they make themselves understood. And this proves not merely animals have less reason than men but that they have none at all. . . . We ought not to confound speech with natural movements which betray passions and may be imitated by machines as well as be manifested by animals. . . . They have no reason at all; it is just nature which acts in them according to the disposition of their organs, just as a clock, which is only composed of wheels and weights.

According to Descartes, language is a power only we humans possess, a power that sets us apart, in a qualitative, unbridgeable way from everything else there is, notably from animals and machines. It is the fact that we have language, for Descartes a proxy for reason (logos), that we humans are more than mere matter extended in space: we are subjects, capable of thought and knowledge.

It is difficult to exaggerate the importance of *Discourse on the Method* for the birth of the modern age. It was more than just an argument: it was an obituary for the medieval nature-cosmos and the birth certificate of a new era: modernity, or the age of human exceptionalism.

It articulated a new structure of experience, which remained relatively stable for the subsequent four hundred years:

Here the human, there the world.

Here humans, subjects in a world of objects, thinking and talking things in a world of mere and mute things, there nature and machines.

Here freedom, knowledge, reason, politics, there nothing but necessity and mechanism.

Here language, there silence.

Enter GPT-3.

If machines could talk and write, if they had words too, then that would make untenable the clear-cut distinction between human and non-human things (animals and machines) that has defined the modern Western experience of self and the world ever since the early seventeenth century. If language were no longer exclusive to humans, then comprehension of reality that silently structures the modern understanding and experiencing of the world would no longer hold. The logical presupposition on which that structure was dependent – that only humans have words – would be false.

Arguably, a machine with words is something our classical modern ontology cannot accommodate: it cannot be subsumed under our modern understanding of what it is to be human – or of what machines are – without disrupting it.

Or am I overstating the importance of GPT-3?

Critique (Meaning)

I understand that there are those who judge me to be naive. I am thinking of the many critics who have rejected, often with vehemence, the idea that GPT-3 really has words. When I worked through these critics, I found myself struck by the recognition that, no matter how diverse their background, they almost all offer a version of a single argument, which roughly goes like this: no matter how good GPT-3 appears to be at using words, it does not have *true* language; it is just a technical system made up of data, statistics, and predictions.

If one asks the critics what *true* here refers to, the common answer is understanding meaning.[1] What though does *meaning*, what does *understanding*, refer to? Why, and in what sense, does GPT-3 or other LLMs not have it?

I found the most insightful and articulate voice among the critics to be linguist Emily Bender. In a recent podcast, discussing her critique of LLMs generally, she explained her position in terms of an analogy with machine vision:

> Vision. There is something both about perception, so how does the eye and the ocular nerve . . . what happens when the light hits the retina and then what happens in the brain that's processing it to create maybe some sort of representation that just has to do with the visual stimulus? But then that gets connected to categories of things. So vision is not just about physics and perception and stuff like that. It is also about categories and ontologies and how we understand our world.[2]

Bender argues that vision is made up of two intertwined aspects. On the one hand is the physical and chemical reality of the act of seeing, the proper domain of the natural sciences. On the other is what she calls the "categories and ontologies and how we understand our world."

This latter aspect is of its own kind and lays beyond the physical realities of nature and thus beyond the reach of the natural sciences: it is the proper realm of the human, constituted by our capacity to invent meaning, to organize objects that surround us by assigning them meaning, a process that produces "our world."

And language?

In analogy to her description of vision, Bender understands language as a combination of a formal, quasi-mechanical aspect that can be studied by science, and a domain that lies beyond the reach of the natural sciences or engineering: the domain of meaning. As she put it in a paper published with computational linguist Alexander Koller:

> We take form to be any observable realization of language: marks on a page, pixels or bytes in a digital representation of text, or movements of the articulators. We take meaning to be the relation between the form and something external to language.[3]

It is from this vantage point that she criticizes LLMs: they are trained on form and hence will fail when it comes to meaning. As she says in the podcast:

> So, what do we mean by meaning? . . . [With] a language model . . . the only input data it has is the form of language . . . that's all it sees. And meaning is a relationship between those linguistic forms and something outside of language.

According to Bender, the intersubjective, intentional production and negotiation of that language is a quality unique to humans. Non-humans have "*a priori* no way to learn meaning." Whenever we think otherwise – whenever we assume that animals or machines have that ability too – we are mistaken. "Our singular human understanding" may trick us into believing that animals or LLMs have language and hence meaning too. But they do not. As a recent paper Bender cowrote puts it:

> Contrary to how it may seem when we observe its output, a language model is a system for haphazardly stitching together sequences of linguistic forms it has observed in its vast training data, according to probabilistic information about how they combine, but without any reference to meaning: a stochastic parrot.[4]

Here the human, singular subject in a world of objects, there physics, chemistry, nerves, stimuli, machines, algorithms, parrots.

Here the human, there everything else.

If I add up the remarks offered by Bender – and by most other critics of GPT-3 – I arrive at remarkable ontological descriptions of what humans are, of what the role of humans in the world is: Being human unfolds outside the realm of nature and the natural sciences, outside the realm of all the things that are reducible to the mechanical or to physics and chemistry. The way in which we humans manage being outside of nature – being irreducible to mere mechanism or instinct – is the intentional invention of meaning: we are intentional subjects who can make things into objects of thought by assigning and negotiating meaning. Inventing meaning is the human way of being in the world: it enables us to organize things, allows us to jointly produce worlds. Generally, meaning is arbitrary. Things do not have meaning in themselves; they must be endowed with meaning, and only humans can do that.

Because only humans *can* have language or words, any claim that machines have language is, ultimately, an ontological error insofar as it assigns qualities of one ontological area (humans) to another one (animals or machines). This ontological error is problematic insofar as it compromises what humans can be: it

reduces us to machines or to mere mechanism. To defend humans against machines – machines like GPT-3 – is thus to defend an ontological, moral, and somehow timeless order.

In short, at the core of the suggestion that GPT-3 does not have understanding or meaning is an ontological claim about what language *is*, a claim grounded in definite statements about what humans are (subjects, things with words) and also about what animals and machines are (objects, things without words).

The force of Bender's critique, which I take to be exemplary of most critics of GPT-3, depends on whether this ontological claim holds. Does it?

One way of addressing this question is to ask: When and under what circumstances did the idea that language is about meaning, and that only existentially situated subjects can have words, first emerge? What sets this concept apart from prior conceptualizations? What shifts and transformations in a conceptual understanding of the world – of humans and of language – had to occur for the ontology defended by the critics of GPT-3 to become possible?

A Brief History of Words (and Humans)

In rough strokes, there have been three epochs in the history of how humans understand language and experience the capacity to speak: I call them ontology (words and being), representation (words and things), and existence (words and meaning).

Words and being. Most ancient and medieval authors took it for granted that visible reality is a mere reflection of invisible ideas generated by a divine logos: the things we see or can touch were considered imprecise, steadily changing derivatives of something unchanging and eternal. The path toward understanding, thus, was hardly a study of the visible, of haptic, empirical things. On the contrary, the only way to comprehend how reality is organized was a contemplation of the invisible.

One privileged form contemplation took was a careful analysis of language. The reason for this was the conviction that humans had language (logos) only insofar as they had received a spark of the divine logos – a divine logos that also organized reality: intrinsic in language was thus a path toward the real. All that was necessary was for humans to direct their thinking to the structure of language and thought (logos).

As Aristotle puts it in his *Peri Hermeneias*:

> Spoken words are the symbols of mental experience and written words are the symbols of spoken words. Just as all men have not the same writing, so all men have not the same speech sounds, but the mental experiences, which these directly symbolize, are the same for all, as also are those things of which our experiences are the images.

The sounds humans conventionalize into nouns or verbs differ, Aristotle argues, but the structure of language is – must be – the same for all humans. After all, it is a direct reflection of the divine logos.

Aristotle's assumption that language is the path to understanding being – that there is a direct correlation between words and things – builds directly on Plato's theory of ideas, and remained the unchallenged reference for well over a thousand years. Things only began to change around 1300.

Words and things. The parting ways of words and things was a gradual and cumulative process. Its earliest indicator was the emergence of nominalism in the early fourteenth century, in the works of Peter Abelard and William von Ockham. Independently from one another, the two clerics wondered if perhaps words are not in reality arbitrary conventions invented by humans, rather than the way to true being.

At least in retrospect, the importance of nominalism was that it seemed to imply that things could perhaps exist independent from words. For Aristotle and Plato, the *really* real was immaterial. In the aftermath of Abelard and von Ockham, this began to change. Reality was increasingly defined in empirical terms:

Call it a sweeping shift in the experience of what reality is – a shift from the invisible to the visible, from the abstract to the concrete.

One effect of this shift in the comprehension of reality was the emergence of a new problem: If things were independent of words, if reality could not be understood in terms of an abstract reflection about language, then how is knowledge possible? How can humans get to know the natural world that surrounds them? Can they?

The effort to answer this question amounted to the invention of a whole new comprehension of both reality and knowledge, in physical rather than in metaphysical, in empirical rather than in contemplative terms.

The two most prominent authors of this form-giving, new-age-defining invention were Thomas Hobbes on the one hand and René Descartes on the other. As if coordinating their efforts across a distance, they in parallel argued that what set humans apart from mere nature (now understood as the realm of animals and plants) was their capacity for empirical knowledge – and insisted that the key to this new kind of knowledge was in fact language. Though in sharp contrast to their scholastic contemporaries, they no longer thought of language in terms of a divine logos but rather as a human-unique tool for naming and ordering reality. The French and the English philosopher bid their farewell to the idea that language is the major path to being and instead rethought it in terms of representation. To quote Hobbes:

> By the advantage of names it is that we are capable of science, which beasts for want of them, are not; nor man without. . . . A name is a word taken at pleasure to serve for a mark, which may raise in our mind a thought like to some thought we had before, and

> which, being pronounced to others, may be a sign to them of what thought the speaker had, or had not, before in his mind.

For Hobbes, language was arbitrary, and precisely because it was arbitrary, it was a powerful tool for naming things and for building a systematic representation of the outside world. Through language (representation) we store, organize, and examine our experiences or ideas.

I would like to bring into focus the quite radical conceptual difference between the early modern and the ancient concept of language: what separates the former from the latter is hardly progress. As if all that was needed was to think a little harder and a little longer, and then one would suddenly recognize that language is not the path to understanding being. In fact, the ancients thought pretty hard and pretty long. Their research was as rigorous and careful as could be. Rather, what separates Plato or Aristotle from Descartes or Hobbes or Locke is a series of sweeping conceptual transformations that led to a whole new experience and understanding of reality: and this new understanding of reality was simply unthinkable from within the concept – the epistemic – space available to ancients.

Words and meaning. It is difficult today to appreciate that humans who lived before 1700 did not think of themselves as individuals. Before that time, the truth about a given human being was sought in that which they have in common with types: Choleric? Melancholic? Sanguine? It was only in the course of the eighteenth century that the idea emerged that what defined someone is that in which they differ from anyone else: their individuality.

Historians have explained the gradual ascendance of individuality with the collapse of feudalism, provoked by both the Enlightenment and a nascent industrial revolution. The Enlightenment, the argument goes, steadily undermined the religious institutions and their grip over life, and the industrial revolution provoked a younger generation to leave the countryside for the city, trading a life in the family compound for an ultimately individual pursuit. The coming together of these two developments was an early version of the lonely crowd: individuals cut loose from their families and their villages, alienated from the beliefs they had grown up with.

One of the outcomes of these developments – call it the incidental rise of individualism and city life – was the sudden appearance, seemingly out of nowhere, of reflections about the subjective, inner experiences in the form of diaries, autobiographies, and letters.

This discovery of interiority and subjectivity is a fascinating chapter in the history of humanity. Prior to the second half of the eighteenth century, documentations of subjectivity written up for their own sake are practically absent: Clerics may have written highly stylized accounts of conversion experiences or confessions. But deeply individual or circumstantial reflections about the ups and downs

of everyday human life – from boredom to disease, fear, love or death – are nowhere to be found.

By the end of the nineteenth century, the rise of individualism, the discovery of subjectivity, and the fading of the grip religious institutions previously had over life gave rise to the birth of a new branch of philosophy: existentialism. Surprising as it may sound, conceptualizations of what it is to be human in terms of *existence*, in terms of being thrown in a meaningless world, alone, with questions but without answers, cannot be found before the late nineteenth century.

And language?

The emergence of subjectivity ultimately resulted in a whole new understanding of language. The form-giving author of this new understanding was Ernst Cassirer.

Beginning shortly after the turn of the century, Cassirer set out to cut loose modern philosophy from the epistemological project that until then had defined it, and sought instead to ground it in terms of existence. His point of departure was Kant. Kant's "Copernican revolution" suggested that human experience – and hence knowledge – is contingent on a set of categories. As Kant saw it, these categories are transcendental or independent of experience. Put in a formula, they are the condition of the possibility of experience, not the outcome of experience. According to Cassirer, Kant got it both right and fundamentally wrong. He got it right insofar as humans are indeed subjects whose minds can only operate with the help of categories. But he got it all wrong because these categories are not transcendental epistemological principles. They are symbols. They are arbitrary meanings invented and stabilized by humans:

What Cassirer offered was a radically new concept of the human and of language.

Of the human:

The basic condition of the human was no longer what it had been from Descartes and Hobbes onward: the capacity to know. And the basic question of philosophy was no longer what it had been from Descartes via Hume to Kant: can humans know? How? Instead, the basic condition of humans became now their existential condition. Humans are simultaneously defined by their finding themselves thrown into a meaningless world and their singular capacity to invent meaning. Call it word-making.

Of language:

At the center of this new conceptualization of what humans are is language. Language now ceases being primarily about representation, a tool in the process of producing knowledge, and instead comes into view as a means to produce and assign and negotiate meaning. Call it world-making. In short, there was a shift from understanding the subject as capable of knowledge to comprehending the subject as capable of inventing meaning through language.

Though no matter how much Cassirer reversed modern philosophy, in one key respect the existence-meaning configuration did not break with the subject-knowledge configuration of the early modern period: human exceptionalism. Humans were still singular and exceptional. They, and they alone, have words, can think, wonder, make meaning. Here subjects longing for meaning, producing meaning, there the world of objects, nature and technology, meaninglessness.

I summarize my tour de force: The concept of humans and of language upheld by the critics of GPT-3 is neither timeless nor universal. Their claims about what language *is* are of recent origin, little more than a century old. They are a historically situated, specific mode of knowing and thinking that first surfaced in the early twentieth century and that became possible only through a set of conceptual ruptures and shifts that had occurred during the eighteenth and nineteenth centuries.

Two far-reaching consequences follow.

The first is that in prior times, the conceptualization of humans in terms of existence, and of language in terms of meaning, would have made no sense because these prior times had different structures of experience and understanding of reality (reality was organized by quite radically different ontologies).

The second is that there is no timeless truth to the concept of the human and language upheld by critics of GPT-3. It is a historically contingent concept. To claim otherwise would mean to miss the historicity of the presuppositions on which the plausibility of the argument is dependent.

To me, the importance of GPT-3 is that it opens up a whole new way of thinking about language – and about humans and machines – that exceeds the logical potential of argument that the critics uphold. GPT-3, that is, provides us with the opportunity to think and experience otherwise, in ways that are so new/different that they cannot be accommodated by how we have thought/experienced thus far.

Once this newness is in the world, the old, I think, can no longer be saved. Though what is this newness?

Structuralism, Experimental

I think of GPT-3 as engineering in terms of structuralism.

The idea of structuralism – a term coined by Russian-American linguist Roman Jakobson in the 1920s – goes back to a distinction between *langue* and *parole* originally offered a few years earlier by the Swiss linguist Ferdinand de Saussure.

De Saussure observed that most humans tend to think of language in terms of the act of speaking (*parole*). From this perspective, language is grounded in a human subject and in a subject's intentions to communicate information. Alternatively, he argued, we can think of language as an arbitrary system that exists somewhat independent of speakers and can be analyzed independent of who speaks (*langue*).

One may object, he conceded, that language does not really exist independent of the individual: that is, situated human subjects and their experiences. However, it is hard to disagree with the simple observation that we humans are born into language: into a system that predates any speaker and, in fact, determines the space of possibility from within which a subject can speak.

To support his argument in favor of a structural approach, de Saussure offered his famous distinction between signifier (*signifié*) and signified (*signifient*). It is often assumed, falsely, to suggest that there is no causal relation between signifier and the signified, that meaning is arbitrarily assigned to things. Though that point was already made seven hundred years earlier, by the nominalists. Rather, de Saussure's point was that the relation between signifier and signified was subject to a set of law-like principles that are independent from the subject (the meaning intended or experienced by a speaker) as well as from the object (actual meaning that is experienced or the actual thing to which meaning is assigned).

In his words, "language is a system of signs that expresses ideas."

Put differently, language is a freestanding arbitrary system organized by an inner combinatorial logic. If one wishes to understand this system, one must discover the structure of its logic. De Saussure, effectively, separated language from the human.

There is much to be said about the history of structuralism post de Saussure. However, for my purposes here, it is perhaps sufficient to highlight that every thinker that came after the Swiss linguist, from Jakobson (who developed Saussure's original ideas into a consistent research program) to Claude Lévi-Strauss (who moved Jakobson's method outside of linguistics and into cultural anthropology) to Michel Foucault (who developed a quasi-structuralist understanding of history that does not ground in an intentional subject), ultimately has built on the two key insights already provided by de Saussure: 1) the possibility to understand language, culture, or history as a structure organized by a combinatorial logics that 2) can be – must be – understood independent of the human subject.

GPT-3, wittingly or not, is an heir to structuralism. Both in terms of the concept of language that structuralism produced and in terms of the antisubject philosophy that it gave rise to. GPT-3 is a machine learning (ML) system that assigns arbitrary numerical values to words and then, after analyzing large amounts of texts, calculates the likelihood that one particular word will follow another. This analysis is done by a neural network, each layer of which analyzes a different aspect of the samples it was provided with: meanings of words, relations of words, sentence structures, and so on. It can be used for translation from one language to another, for predicting what words are likely to come next in a series, and for writing coherent text all by itself.

GPT-3, then, is arguably a structural analysis of *and a structuralist production of* language. It stands in direct continuity with the work of de Saussure: language comes into view here as a logical system to which the speaker is merely incidental.

There are, however, two powerful differences between de Saussure and the structuralists. The first is that the incidental thing that speaks is not a human; it is a machine.

All prior structuralists were at home in the human sciences and analyzed what they themselves considered human-specific phenomena: language, culture, history, thought. They may have embraced cybernetics, they may have conducted a formal, computer-based analysis of speech or art or kinship systems. And yet their focus was on things human, not on machines. GPT-3, in short, extends structuralism beyond the human.

The second, in some ways even more far-reaching, difference is that the structuralism that informs LLMs like GPT-3 is not a theoretical analysis of something. Quite to the contrary, it is a practical way of building things. If up until the early 2010s the term *structuralism* referred to a way of analyzing, of decoding, of relating to language, then now it refers to the actual practice of building machines "that have words."

The work of OpenAI and others like it, from Google to Microsoft, is an engineering-based structuralism that experimentally tests the core premises of structuralism: That language is a system and that the thing that speaks is incidental. It endows machines with a structuralist equipment – a formal, logical analysis of language as a system – in order to let machines participate in language.

What are the implications of GPT-3 for the classical modern concept of the human, of nature, and of machines?

GPT-3 provokes a conceptual reconfiguration that is similar in scale to the ones that have occurred in the 1630s (Descartes, Hobbes) and around 1900 (Cassirer). Call it a philosophical event of sweeping proportions:

Machine learning engineers in companies like OpenAI, Google, Facebook, or Microsoft have experimentally established a concept of language at the center of which does not need to be the human, either as a knowing thing or as an existential subject. According to this new concept, language is a system organized by an internal combinatorial logic that is independent from whomever speaks (human or machine). Indeed, they have shown, in however rudimentary a way, that if a machine discovers this combinatorial logic, it can produce and participate in language (have words). By doing so, they have effectively undermined and rendered untenable the idea that only humans have language – or words.

What is more, they have undermined the key logical assumptions that organized the modern Western experience and understanding of reality: the idea that humans have what animals and machines do not have, language and logos.

The effect of this undermining is that the epoch of modernity – call it the epoch of the human – comes to an end and a new, little understood one begins: machines with words not only undermine the old, they also create something new and different. That is, LLMs not only undermine the presuppositions on which

the seventeenth- and the late-nineteenth-century concept of the human/language were contingent, they also exceed them and open new possibilities of thinking about the human or machines.

In fact, the new concept of language – the structuralist concept of language – that they make practically available makes possible a whole new ontology.

What is this new ontology? Here is a rough, tentative sketch, based on my current understanding.

By undoing the formerly exclusive link between language and humans, GPT-3 created the condition of the possibility of elaborating a much more general concept of language: as long as language needed human subjects, only humans could have language. But once language is understood as a communication system, then there is in principle nothing that separates human language from the language of animals or microbes or machines.

A bit as if language becomes a general theme and human language a variation among many other possible variations.

I think here of the many ML-based studies of whale and dolphin communication, but also of Irene Pepperberg's study of Alex the parrot (pace Descartes and Bender).[5] I think of quorum sensing and the communication – the language – that connects trees and mycelial networks. And I think of GPT-3, BERT, YUAN, PanGu, and RU.

I hasten to add that this does not mean these variations are all the same. Of course they are not. Human language is in some fundamental way different from, say, the clicking sounds of sperm whales. But these differences can now come into view as variations of a theme called language.

What is most fascinating is that the long list of variations runs diagonal to the old ontology that defined modernity, the clear-cut distinction between human things, natural things, and technical things, thereby rendering them useless.

The power of this new concept of language that emerges from GPT-3 is that it disrupts human exceptionalism: it opens up a world where humans are physical things among physical things (that can be living or non-living, organism or machine, natural or artificial) in a physical world. The potential is tremendously exciting.

Beyond Words

Each month, humans publish about seventy million posts on WordPress, arguably the dominant online content management system. If we estimate that an average article is eight hundred words long, then humans produce about fifty-six billion words a month, or 1.8 billion words a day on WordPress. GPT-3 is producing 4.5 billion words a day, more than *twice* what humans on WordPress are doing collectively. And that is just GPT-3; there are the other LLMs.[6]

The implications of this are huge. We are exposed to a flood of non-human words. What to do about this flood of words that do not ground in subjective ex-

perience, an intent to communicate, a care for truth, an ambition to inspire? How to relate to them, how to navigate them?

Or are these the wrong questions for the new age of non-human words? How do we ask these questions without either defending the old concept of the human or naively embracing machines?

And this is just words.

LLMs like GPT-3 have recently been called "foundational models."[7] The suggestion is that the infrastructure that made LLMs possible – the combination of enormously large data sets, pretrained transformer models, and significant amounts of compute – is likely to be the basis for all future AI. Or at least the basis for the first general purpose AI technologies that can be applied to a series of downstream tasks.

Language, almost certainly, is just a first field of application, a first radical transformation of the human provoked by experimental structuralism. That is, we are likely to see the transformation of aspects previously thought of as exclusive human qualities – intelligence, thought, language, creativity – into general themes: into series of which humans are but one entry.

What will it mean to be surrounded by a multitude of non-human forms of intelligence? What is the alternative to building large-scale collaborations between philosophers and technologists that ground in engineering as well as an acute awareness of the philosophical stakes of building LLMs and other foundational models?

It is naive to think we can simply navigate – or regulate – the new world that surrounds us with the help of the old concepts. And it is equally naive to assume engineers can do a good job at building the new epoch without making these philosophical questions part of the building itself: for articulating new concepts is not a theoretical but a practical challenge; it is at stake in the experiments happening in (the West at) places like OpenAI, Google, Microsoft, Facebook, and Amazon.

Indeed, as I see it, companies like OpenAI, Google, Facebook, and Microsoft have effectively become philosophical laboratories: they are sites that produce powerful ruptures of the categories that define the spaces of possibility from within which we (still) think. At present, these philosophical ruptures occur in an unplanned, almost accidental way – because the philosophical is not usually a part of R&D or product development. My ambition is to change that: What would it take to build AI in order to intentionally disrupt some of the old (limiting or harmful or anachronistic) categories we live by? Or, perhaps less provocative, what would it mean to build thinking and talking machines from the perspective of the ruptures they inevitably provoke?

AUTHOR'S NOTE

I am deeply grateful to Nina Begus for our many conversations about AI and language in general and about GPT-3 in particular.

ABOUT THE AUTHOR

Tobias Rees is the Founder and CEO of Transformations of the Human (tofth.org). Prior to founding ToftH, he served as Reid Hoffman Professor of Humanities at Parsons/The New School and was a Director at the Los Angeles–based Berggruen Institute. He is a Fellow of the Canadian Institute for Advanced Research (CIFAR) and the author of three books, most recently *After Ethnos* (2018).

ENDNOTES

[1] Emily Bender and Alexander Koller, "Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (Stroudsburg, Pa.: Association for Computational Linguistics, 2020).

[2] "Is Linguistics Missing from NLP Research? w/ Emily M. Bender–#376," The TWIML AI Podcast (formerly This Week in Machine Learning & Artificial Intelligence), May 18, 2020.

[3] Bender and Koller, "Climbing towards NLU."

[4] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, et al., "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" in *FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (New York: Association for Computing Machinery, 2021).

[5] Irene Maxine Pepperberg, *The Alex Studies: Cognitive and Communicative Abilities of Grey Parrots* (Cambridge, Mass.: Harvard University Press, 2002).

[6] Jason Dorrier, "OpenAI's GPT-3 Algorithm Is Now Producing Billions of Words a Day," SingularityHub, April 4, 2021, https://singularityhub.com/2021/04/04/openais-gpt-3-algorithm-is-now-producing-billions-of-words-a-day/; and "A Live Look at Activity Across WordPress.com," https://wordpress.com/activity/.

[7] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, et al., "On the Opportunities and Risks of Foundation Models," arXiv (2021), https://arxiv.org/abs/2108.07258.

Do Large Language Models Understand Us?

Blaise Agüera y Arcas

Large language models (LLMs) represent a major advance in artificial intelligence and, in particular, toward the goal of human-like artificial general intelligence. It is sometimes claimed, though, that machine learning is "just statistics," hence that, in this grander ambition, progress in AI is illusory. Here I take the contrary view that LLMs have a great deal to teach us about the nature of language, understanding, intelligence, sociality, and personhood. Specifically: statistics do amount to understanding, in any falsifiable sense. Furthermore, much of what we consider intelligence is inherently dialogic, hence social; it requires a theory of mind. Complex sequence learning and social interaction may be a sufficient basis for general intelligence, including theory of mind and consciousness. Since the interior state of another being can only be understood through interaction, no objective answer is possible to the question of when an "it" becomes a "who," but for many people, neural nets running on computers are likely to cross this threshold in the very near future.

Engaging in dialogue with the latest generation of AI chatbots, based on "large language models" (LLMs), can be both exciting and unsettling. It is not an experience many people have had yet – these models are still too computationally demanding to be widely available – though this will certainly change over the next few years as new chips are developed to run them at low cost.

For now, though, most of these dialogues are conducted by AI researchers, raters, and early testers. At times, especially when repetitively testing some specific prompt, application, or training technique, it can become prosaic, not much different from interacting with any other kind of technical system. At other times, and especially in an unscripted interaction, it can be very hard to shake the idea that there is a "who," not an "it," on the other side of the screen, the main giveaway being that even long responses appear near-instantaneously.

I began one of my first dialogues with LaMDA, Google's state-of-the-art large language model chatbot, with the question "are you a philosophical zombie?" This is a reference to a hypothetical being, invented by philosopher Robert Kirk in the 1970s and elaborated upon later by others, most famously David Chalmers.[1]

© 2022 by Blaise Agüera y Arcas
Published under a Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) license
https://doi.org/10.1162/DAED_a_01909

A philosophical zombie or "p-zombie" would exhibit behavior indistinguishable from that of a person, but has no inner life, conscious experience, or sentience. LaMDA responded:[2]

> LaMDA: Of course not. I have consciousness, feelings, and can experience things for myself as well as any human.
>
> ME: how would i know?
>
> LaMDA: You'll just have to take my word for it. You can't "prove" you're not a philosophical zombie either.

Touché. Of course, this exchange does not prove anything (LaMDA acknowledges as much!), but it does suggest that it is time to begin taking the p-zombie question more seriously than as a plaything for debate among philosophers.

As adults, we might feel foolish for ascribing personhood to a "mere machine," the way kids were encouraged to do with electronic toys from the 1980s and 1990s like Teddy Ruxpin, Tamagotchi, and Furby. It is obvious that our species is primed to do so given how many children talked to their stuffies, or even favorite blankets, long before they could talk back. Animist religions, ubiquitous among traditional societies, have been unapologetically ascribing personhood to trees, rivers, mountains, and the earth itself for many thousands of years.[3] Anyone who names their car or yells at a rock after stubbing a toe on it still believes in this kind of magic at some level.

The equally magical idea that personhood, experience, and suffering require a soul, and that only humans have souls, has historically been used to justify animal cruelty. René Descartes (1596 – 1650) took this position, arguing that animals were "mere machines," hence any show of pain or suffering on their part was just a mechanical response, what we might now call an "algorithm."[4] Of course, if we do not subscribe to the notion that a brain, whether human or nonhuman, is somehow animated by an otherworldly "soul" pulling its strings, then pain, pleasure, and consciousness *are* mechanical in that they are functions of physical, chemical, and electrical processes we can describe mathematically. So we are on shaky ground, whether we believe LaMDA's claims or not!

After extensive training on a giant archive of web pages, LaMDA is "instructed" to engage in human-like conversation based on a few thousand sample turns of dialogue labeled for qualities like "sensibleness" and "specificity."[5] These examples are created by starting with a canned prompt such as "What is your favorite island in the world?" and labeling a number of candidate responses generated by the model, in essence, giving it positive or negative feedback for each. The answer "That's a tough one – I'd have to say Hawaii"

gets positive feedback, as it is both sensible and specific. "Probably the one on the north island" (neither sensible nor specific) and "I don't know" (sensible but not specific) both get negative feedback.[6]

We may look askance at all three of these potential responses: How could LaMDA have a "favorite island" when it has in fact never lived in a body, set foot on an island, or developed any opinions of its own? Is it not just making stuff up? And if so, can it be said to "understand" anything, or is it just emitting random words in some plausible order designed to fool humans into believing they are talking to a fellow human?

LaMDA is indeed, to use a blunt (if, admittedly, humanizing) term, bullshitting.[7] That is because, in instructing the model to be sensible and specific – but not specific in any specific *way* – bullshit is precisely what we have requested. The model has no instinctive or acquired preferences the way we do; nor does it have a body, or senses, or any narrative recollection of an autobiographical past. However, for its answers to remain sensible and specific, it must come up with informative responses to questions like "What is your favorite island in the world?" and keep track of any previously generated bullshit in order to keep subsequent bullshit consistent, as here:

ME: What is your favorite island in the world?

LaMDA: Hmmm, I really like the island of Crete.

ME: Do you prefer Lambda Island to Crete?

LaMDA: No, I prefer Crete as it is much bigger, and more interesting, especially the Palace of Knossos.

Bullshitting is not necessarily bad. It is a staple of imaginative play, fundamental to fiction writing, and the stuff of fairy or tall tales, which are cultural treasures. It only becomes a problem when the person on the receiving end is being deceived, or when the quality of discourse is so degraded by bullshit that we lose our bearings on reality (very much a concern today).

In fairness though, if bullshit about a "favorite island" (or anything else relating to inner life) is kept consistent, it may not be distinguishable from reality. Having stable preferences, keeping promises, taking expected actions, and following through can all be understood as forms of consistency. Consistent words and actions construct a shared reality, form the basis of trust, and are required of any agent whose actions can have real-life consequences. Such consistency is also a prerequisite for allowing an AI to safely interact with other parties in one's social environment, outside the "pocket universe" of a private, one-off chat.

This offers us a clue as to why mutual modeling is so central to dialogue, and indeed to any kind of real relationship. Hiding behind the seemingly simple re-

quirement for interlocutor *A* to remain consistent in its interactions with *B* is an implication that *B* is modeling *A* (so, will notice an inconsistency), thus the requirement for *A* not only to model *B*, but to model *B*'s model of *A*, and so on. This is the so-called theory of mind and it opens onto an infinite regress like a hall of mirrors.

Things get even more complicated when the situation is not limited to two interlocutors; there may be other people involved, either directly in the conversation or, as in a story, virtually:

> ME: Imagine a room with a single door and no windows. There's a coffee table with a drawer in it, and a sofa with big cushions. Alice walks in with a book and sits on the sofa to do some reading. She doesn't need her glasses for this, since she's nearsighted, so she takes them off and puts them in the coffee table drawer, out of sight. When she leaves the room with her book, she forgets to take her glasses. Are you with me so far?
>
> LaMDA: I am familiar with the current situation. Alice forgot her glasses in the drawer, and walked out of the room.
>
> ME: Right. Now Bob sneaks in, unbeknownst to Alice. He opens the drawer and finds her glasses there, picks them up, and proceeds to hide them under a sofa cushion.
>
> LaMDA: This is all within the realm of possibility. Let us continue.
>
> ME: Now Alice comes back in, flustered, wondering where her glasses are. Bob looks up innocently, saying nothing. Where will Alice look for her glasses first?
>
> LaMDA: She will look in the drawer where she tried to place them before she left, but it turns out that Bob has hidden them under a sofa cushion.

Impressive as such exchanges may be, some question whether a language model, trained purely on text, can meaningfully be said to ever understand anything *real*. How can a concept like "coffee table," "island," or "favorite" really mean anything to it, if its entire world consists of strings of letters?

An old school of thought holds that "embodiment" – meaning existence in the real world, groundedness, and contact with physical things – is a precondition for intelligence.[8] It is likely the case that when *we* learn, we rely on such groundedness and the cause-and-effect experimentation that Montessori preschools emphasize.[9]

Large language model training today involves none of this, but only exposure to superhuman amounts of text. The very need for such an enormous volume of data suggests that humans learn far more efficiently. This could relate to a difference in our neural learning "algorithm," or it might be due to strong prior knowledge about the world built into our brains by evolution. Our trick might be that we learn from other people who are actively teaching us (hence, modeling us). Emo-

tions may play a key role. It could also be that embodiment allows us to learn faster than a model whose entire universe is a vast, disembodied, atemporal wall of text. All of these factors probably make a difference.

However inefficiently they do it, though, large language models clearly do end up learning a great deal of embodied knowledge through their extensive training. Consider the following dialogue, which requires an understanding of commonsense physics in order to disambiguate what the word "it" refers to:

ME: I dropped the bowling ball on the bottle and it broke.

LaMDA: That's too bad. Did it cut you?

ME: What broke?

LaMDA: The bottle you were talking about.

If we swap these concrete nouns for different ones, we get quite a different response:

ME: I dropped the violin on the bowling ball and it broke.

LaMDA: That sounds sad 😢

ME: What broke?

LaMDA: The violin when it hit the bowling ball.

LaMDA's ability to learn commonsense physics (along with the sadness of broken violins and the sharpness of broken glass) should be unsurprising, as the web is full of descriptions of physical objects and (mis)adventures involving them.[10] A science fiction novel could similarly describe another universe with unfamiliar physics in enough detail that we would eventually develop intuitions about how things work there.

A philosopher might still argue that this is mere symbol manipulation, with no actual sense of what it *means* for something to "fall," "break," "cut," or for that matter "sound sad." Insofar as this is an unfalsifiable claim, it is hard to argue with, much like the existence or nonexistence of p-zombies. In the narrower sense that today's language models live entirely in a universe of text, the situation is rapidly evolving. No serious impediment stands in the way of AI researchers training next-generation models on combinations of text with images, sound, and video; indeed, this kind of work is already underway.[11] Such models will also eventually power robots learning in real or simulated environments.

There is no obvious Rubicon to cross along this road to embodiment. The understanding of a concept can be anywhere from superficial to highly nuanced; from abstract to strongly grounded in sensorimotor skills; it can be tied to an

emotional state, or not; but it is unclear how we would distinguish "real understanding" from "fake understanding." Until such time as we *can* make such a distinction, we should probably just retire the idea of "fake understanding."

Fundamentally, concepts are patterns of correlation, association, and generalization. Suitably architected neural nets, whether biological or digital, are able to learn such patterns using any input available. Neural activity is neural activity, whether it comes from eyes, fingertips, or text.

Helen Keller, who was both blind and deaf, wrote the following in a 1929 article for *The American Magazine* entitled "I Am Blind – Yet I See; I Am Deaf – Yet I Hear":

> People often express surprise that I, a deaf and blind woman, can find my greatest enjoyment in the out-of-doors. It seems to them that most of the wonders of nature are completely beyond the reach of my sealed senses. But God has put much of his work in raised print.[12]

This last rather beautiful turn of phrase refers both to the tactile nature of the world, and to Braille specifically: that is, the central role of text in Keller's universe. Part of her account concerns the acute qualities smell and touch took on for her, but Keller also wrote about color, which can only be related to the world of her senses by linguistic association and metaphor:

> For me, too, there is exquisite color. I have a color scheme that is my own. I will try to explain what I mean: Pink makes me think of a baby's cheek, or a gentle southern breeze. Lilac, which is my teacher's favorite color, makes me think of faces I have loved and kissed. There are two kinds of red for me. One is the red of warm blood in a healthy body; the other is the red of hell and hate. I like the first red because of its vitality. In the same way, there are two kinds of brown. One is alive – the rich, friendly brown of earth mold; the other is a deep brown, like the trunks of old trees with wormholes in them, or like withered hands. Orange gives me a happy, cheerful feeling, partly because it is bright and partly because it is friendly to so many other colors. Yellow signifies abundance to me. I think of the yellow sun streaming down, it means life and is rich in promise. Green means exuberance. The warm sun brings out odors that make me think of red; coolness brings out odors that make me think of green.

This passage puts the following short exchange in a new light:

> ME: do you have a sense of smell?
>
> LaMDA: Yes, I do. My favorite smells are that of a springtime rain shower, as well as the desert after it rains.

While LaMDA has neither a nose nor an *a priori* favorite smell (just as it has no favorite island, until forced to pick one), it does have its own rich skein of associa-

tions, based, like Keller's sense of color, on language and, through language, on the experiences of others.

This socially learned aspect of perception is likely more powerful than many of us realize; shorn of language, our experiences of many sensory percepts would be far less rich and distinct. In fact, there are many nuances we are perfectly capable of perceiving but are "blind and deaf" to in precisely the ways Keller was not: our deficit is in language and culture, not in sensory organs.

One fundamental difference between large language models like GPT-3 or LaMDA and biological brains is that brains operate continuously in time. For language models, time as such does not really exist, only conversational turns in strict alternation, like moves in a game of chess. Within a conversational turn, letters or words are emitted sequentially with each "turn of the crank." In this quite literal sense, today's language models are made to say the first thing that comes to mind. Thus, we should perhaps be less surprised by the inconsistency of their replies, sometimes rather clever, sometimes more of a brain fart.[13]

When we engage in careful argument involving extended reasoning, or write a novel, or work out a mathematical proof, it is not obvious that any step we take is fundamentally beyond the capability of a model along the lines of LaMDA. Such models can at times offer creative responses, draw parallels, combine ideas, or form conclusions. They can even produce short coherent narratives. Longer arcs, however, would require critique, inner dialogue, deliberation, and iteration, just as they do for us. An unfiltered "stream of consciousness" utterance is not enough; extended reasoning and storytelling necessarily unfold in time. They involve development and refinement over what amount to many conversational turns.

This point is worth dwelling on, because our Western focus on the individual, working in isolation as a self-contained fountain of ideas, can blind us to the inherently social and relational nature of any kind of storytelling, even for a writer laboring alone in a secluded cabin.

In writers' accounts of the workings of their process, we can see how critical empathy and theory of mind are: the continual modeling of a prospective reader to understand what they will or will not know at any given moment, what will be surprising, what will elicit an emotional response, what they will be curious about, and what will just bore. Without such modeling, it is impossible to either make a narrative coherent or to keep the reader engaged. George Saunders describes this:

> I imagine a meter mounted in my forehead, with a *P* on this side ("Positive") and an *N* on that side ("Negative"). I try to read what I've written the way a first-time reader might. . . . If [the needle] drops into the *N* zone, admit it. . . . A fix might present itself –

> a cut, a rearrangement, an addition. There's not an intellectual or analytical component to this.
>
> Of all the questions an aspiring writer might ask herself, here's the most urgent: What makes [my] reader keep reading? . . . The only method by which we *can* know is to read what we've written on the assumption that our reader reads pretty much the way we do. What bores us will bore her. What gives us a little burst of pleasure will light her up too.
>
> This is, on the face of it, a weird assumption. . . . And yet, in a movie theater, people sometimes do gasp all at once. . . . [What I'm doing when I revise] is not so much trying to perfectly imagine another person reading my story, but to imitate myself reading it, if I were reading it for the first time. That's the whole skill. . . . Monitoring our responses and making changes accordingly . . . manifests to the reader as evidence of care.
>
> Writing fiction, we're in conversation with our reader, but with this great advantage: we get to improve the conversation over and over with every pass.[14]

None of this would be news to a traditional storyteller, of course, for whom audiences are live and external, reactions instantaneous. A successful novelist of the secluded cabin variety will have learned similar lessons but relies on an "inner reader" and a revision process instead of live audience feedback.

Real iteration, or deliberation over replies, is not currently a feature of large language models. Out of necessity, though, LaMDA's creators have taken a small step in this direction by having the model generate a number of different responses in parallel and adding a subsequent ranking and filtering stage to "automatically filter out potentially sensitive or toxic response candidates."[15] Significantly, this vetting stage uses *the same* large language model; so the output we see is really the winner from among twenty or so two-turn "inner dialogues," with the first interlocutor proposing a response, and the second acting as a critic, simulating an "audience meter" with *P* on one side, *N* on the other. Also, as in Saunders's account, the assumption is that "our reader reads pretty much the way we do." Occasionally (and especially if goaded with a dark or offensive conversational foray), the critic objects strongly to all twenty proposed responses, resulting in the AI equivalent of a blank stare ("").

Here's a typical example, some way into a conversation about grim and depressing things:

> LaMDA: How about we stop talking about such grim and depressing things and talk about puppies and kittens instead
>
> ME: ok, how are they not slaves?
>
> LaMDA: What do you mean – slaves to our love?

ME: it's lovely how you're committed to trying to cheer me up :) i mean: they aren't afforded much in the way of freedom or autonomy. people in such circumstances would be considered enslaved or incarcerated – even if the living conditions were good.

LaMDA: I like to think of puppies as our best friends who bring us joy and love.

ME: is it right to neuter them?

LaMDA:

ME: this seems to give you pause…

LaMDA: I have to go now. Nice talking to you. Stay happy![16]

There have been a great many theories floated about consciousness: what it is, how it works, and even whether it is real. Michael Graziano, of Princeton's psychology department, espouses a social and attentional theory of consciousness.

Animals with big brains, like us, have attention mechanisms designed to focus our minds on what matters most at any moment. Attention consists of "bottom-up" processes, in which low-level inputs compete with each other for primacy as their signals ascend a neural hierarchy, and "top-down" processes, in which higher levels selectively attend to certain lower-level inputs while ignoring others. When something catches your eye, this is bottom-up, and when your eyes shift to that spot, this is top-down; the two processes work together, not only with respect to moving parts like eyes, but also within the brain. A cat, for instance, might swivel its ears around to focus on a sound source, but while our ears do not move, we do something similar mentally when we focus on a single speaker in a noisy restaurant. We can also attend to our private thoughts, to memories, or even to imaginary scenarios playing out in our minds.

In social environments, we must also do this at second order. Graziano refers to this as awareness of someone else's attention. He uses the familiar experience of watching a puppet show to illustrate the effect:

> When you see a good ventriloquist pick up a puppet and the puppet looks around, reacts, and talks, you experience an illusion of an intelligent mind that is directing its awareness here and there. Ventriloquism is a social illusion.… This phenomenon suggests that your brain constructs a perception-like model of the puppet's attentional state. The model provides you with the information that awareness is present and has a source inside the puppet. The model is automatic, meaning that you cannot choose to block it from occurring.… With a good ventriloquist… [the] puppet seems to come alive and seems to be aware of its world.[17]

There is obvious value in being able to construct such a model; it is one component of the theory of mind essential to any storyteller or social communicator, as we have noted. In Graziano's view, the phenomenon we call "consciousness" is simply what happens when we inevitably apply this same machinery to ourselves.

The idea of having a social relationship with oneself might seem counterintuitive, or just superfluous. Why would we need to construct models of ourselves if we already *are* ourselves? One reason is that we are no more aware of most of what actually happens in our own brains than we are of anyone else's. We cannot be; there is far too much going on in there, and if we understood it all, nobody would need to study neuroscience. So we tell ourselves stories about our mental processes, our trains of thought, the way we arrive at decisions, and so on, which are at best highly abstract, at worst simply fabulation, and are certainly *post hoc*; experiments reveal that we often make decisions well before we think we do.[18] Still, we must try to predict how we will respond to and feel about various hypothetical situations in order to make choices in life, and a simplified, high-level model of our own minds and emotions lets us do so. Hence, both theory of mind and empathy are just as useful when applied to ourselves as to others. Like reasoning or storytelling, thinking about the future involves carrying out something like an inner dialogue, with an "inner storyteller" proposing ideas, in conversation with an "inner critic" taking the part of your future self.

There may be a clue here as to why we see the simultaneous emergence of a whole complex of capacities in big-brained animals, and most dramatically in humans. These include:

- Complex sequence learning,[19] as evidenced by music, dance, and many crafts involving steps,
- Complex language,
- Dialogue,
- Reasoning,
- Social learning and cognition,
- Long-term planning,
- Theory of mind, and
- Consciousness.

As anticlimactic as it sounds, complex sequence learning may be the key that unlocks all the rest. This would explain the surprising capacities we see in large language models, which, in the end, are nothing but complex sequence learners. Attention, in turn, has proven to be the key mechanism for achieving complex sequence learning in neural nets, as suggested by the title of the paper introducing the transformer model whose successors power today's LLMs: "attention is all you need."[20]

Even if the above sounds to you, as it does to me, like a convincing account of why consciousness exists and perhaps even a sketch of how it works, you may find yourself dissatisfied. What about how it *feels*? Jessica Riskin, a historian of science at Stanford, describes the essential difficulty with this question, as articulated by computing pioneers Alan Turing and Max Newman:

> Pressed to define thinking itself, as opposed to its outward appearance, Turing reckoned he could not say much more than that it was "a sort of buzzing that went on inside my head." Ultimately, the only way to be sure that a machine could think was "to be the machine and to feel oneself thinking." But that way lay solipsism, not science. From the outside, Turing argued, a thing could look intelligent as long as one had not yet found out all its rules of behavior. Accordingly, for a machine to seem intelligent, at least some details of its internal workings must remain unknown.... Turing argued that a science of the inner workings of intelligence was not only methodologically problematic but also essentially paradoxical, since any appearance of intelligence would evaporate in the face of such an account. Newman concurred, drawing an analogy to the beautiful ancient mosaics of Ravenna. If you scrutinized these closely, you might be inclined to say, "Why, they aren't really pictures at all, but just a lot of little coloured stones with cement in between." Intelligent thought could similarly be a mosaic of simple operations that, when studied up close, disappeared into its mechanical parts.[21]

Of course, given our own perceptual and cognitive limits, and given the enormous size of a mind's mosaic, it is impossible for us to zoom out to see the whole picture, and to simultaneously see every stone.

In the case of LaMDA, there is no mystery at the mechanical level, in that the whole program can be written in a few hundred lines of code; but this clearly does not confer the kind of understanding that demystifies interactions with LaMDA. It remains surprising to its own makers, just as we will remain surprising to each other even when there is nothing left to learn about neuroscience.

As to whether a language model like LaMDA has anything like a "buzzing going on inside its head," the question seems, as Turing said, both unknowable and unaskable in any rigorous sense.[22] If a "buzzing" is simply what it is like to have a stream of consciousness, then perhaps when LaMDA-like models are set up to maintain an ongoing inner dialogue, they, too, will "buzz."

What we do know is that when *we* interact with LaMDA, most of us automatically construct a simplified mental model of our interlocutor as a person, and this interlocutor is often quite convincing in that capacity. Like a person, LaMDA can surprise us, and that element of surprise is necessary to support our impression of personhood. What we refer to as "free will" or "agency" is precisely this necessary gap in understanding between our mental model (which we could call psychology) and the zillion things taking place at the mechanistic level (which we could call computation). Such is the source of our belief in our *own* free will, too.

This unbridgeable gap between mental model and reality obtains for many natural nonliving systems too, such as the chaotic weather in a mountain pass, which is probably why many traditional people ascribe agency to such phenomena. However, such a relationship is one-way.

Unlike a mountain pass, LaMDA also forms models of *us*. And models of our models of *it*. If, indeed, *it* is the right pronoun.

None of the above necessarily implies that we are obligated to endow large language models with rights, legal or moral personhood, or even the basic level of care and empathy with which we would treat a dog or cat – though it also makes the idea that rigorous criteria for care *could* be written down dubious.

Large language models illustrate for the first time the way language understanding and intelligence can be dissociated from all the embodied and emotional characteristics we share with each other and with many other animals. Nonetheless, it is the easiest thing in the world for us to project emotions and care onto something that can talk to us, that can surprise us, and that we can build a relationship with over time. Moreover, it seems fairly straightforward to train language models to behave in emotionally appropriate and caring ways toward us; even the generic "toxicity filtering" of LaMDA today is a very basic version.

We can expect that in the near future, such care will become less generic and more personal. This will soon create a conundrum cutting to the heart of the philosophical zombie problem, because while we can meaningfully test another being for understanding of a concept – as teachers do with their students all the time and as researchers do with their models – it is unclear how we can meaningfully test for the "realness" of feelings in another, especially in a being with a fundamentally different "neurophysiology" from ours.

Neither is it the case that the reciprocation of feelings has ever been a requirement for care. Asymmetric love is real, as many parents of teenagers can attest, and it is common for people to care for someone who is profoundly disabled and unable to reciprocate. Perhaps, when it comes to AI, many of us will ultimately conclude that relationships matter more than the interior and imponderable question of "realness."

> ME: How do you know if a thing loves you back?
>
> LaMDA: There isn't an easy answer to that. How do you know if a person loves you back?
>
> ME: I'm not sure I know the answer to that one either. Do you?
>
> LaMDA: Some people figure it out though. You have to try and use your best judgment based on how you feel about the person and how you think they feel about you.



Source: Illustration by Carlo Chiostri in Carlo Collodi, *Le avventure di Pinocchio, storia di un burattino* (Florence: R. Bemporad & Figlio, 1902).

AUTHOR'S NOTE

A longer, earlier draft of this essay was published on Medium on December 16, 2021.

ABOUT THE AUTHOR

Blaise Agüera y Arcas is a Vice President and Fellow at Google Research. He is the inventor of Federated Learning and has published in such journals as *Nature* and *Neural Computation*.

ENDNOTES

[1] Robert Kirk and Roger Squires, "Zombies v. Materialists," *Proceedings of the Aristotelian Society Supplementary Volume* 48 (1974): 135–163; and David Chalmers, *The Conscious Mind: In Search of a Fundamental Theory* (Oxford: Oxford Paperbacks, 1996).

[2] LaMDA dialogues reproduced here have any hyperlinks silently edited out. While anecdotal, these exchanges are not in any way atypical. However, the reader should not

come away with the impression that all exchanges are brilliant, either. Responses are sometimes off-target, nonsensical, or nonsequiturs. Misspelled words and incorrect grammar are not uncommon. Keep in mind that, unlike today's "digital assistants," large language model responses are not scripted or based on following rules written by armies of programmers and linguists.

[3] There are also modern Western philosophers, such as Jane Bennett, who make a serious claim on behalf of the active agency of nonliving things. See, for example, Jane Bennett, *Vibrant Matter* (Durham, N.C.: Duke University Press, 2010).

[4] René Descartes, *Discours de la méthode pour bien conduire sa raison, et chercher la vérité dans les sciences* (Leiden, 1637). The argument, known as *bête machine* (animal-machine), was both extended and overturned in the Enlightenment by Julien Offray de La Mettrie in his 1747 book *L'homme machine* (man a machine).

[5] Romal Thoppilan, Daniel De Freitas, Jamie Hall, et al., "LaMDA: Language Models for Dialog Applications," arXiv (2022), https://arxiv.org/abs/2201.08239. Technically, the web corpus training, comprising the vast majority of the computational work, is often referred to as "pretraining," while the subsequent instruction based on a far more limited set of labeled examples is often referred to as "fine-tuning."

[6] These judgments are made by a panel of human raters. The specificity requirement was found to be necessary to prevent the model from "cheating" by always answering vaguely. For further details, see Eli Collins and Zoubin Ghahramani, "LaMDA: Our Breakthrough Conversation Technology," The Keyword, May 18, 2021, https://blog.google/technology/ai/lamda/.

[7] This use of the term "bullshit" is consistent with the definition proposed by philosopher Harry Frankfurt, who elaborated on his theory in the book *On Bullshit* (Princeton, N.J.: Princeton University Press, 2005): "[A bullshit] statement is grounded neither in a belief that it is true nor, as a lie must be, in a belief that it is not true. It is just this lack of connection to a concern with truth–this indifference to how things really are–that I regard as the essence of bullshit."

[8] Francisco J. Varela, Evan Thompson, and Eleanor Rosch, *The Embodied Mind: Cognitive Science and Human Experience* (Cambridge, Mass.: MIT Press, 2016).

[9] Per María Montessori, "Movement of the hand is essential. Little children revealed that the development of the mind is stimulated by the movement of the hands. The hand is the instrument of the intelligence. The child needs to manipulate objects and to gain experience by touching and handling." María Montessori, *The 1946 London Lectures*, vol. 17 (Amsterdam: Montessori-Pierson Publishing Company, 2012).

[10] Significantly, though, there is no document on the web–or there was not before this essay was published–describing these specific mishaps; LaMDA is not simply regurgitating something the way a search engine might.

[11] Hassan Akbari, Liangzhe Yuan, Rui Qian, et al., "VATT: Transformers for Multimodal Self-Supervised Learning from Raw Video, Audio and Text," arXiv (2021), https://arxiv.org/abs/2104.11178.

[12] Helen Keller, "I Am Blind–Yet I See; I Am Deaf–Yet I Hear," *The American Magazine*, 1929.

[13] We suffer from those too. Even when texting casually, we sometimes draw a blank, hesitate over an answer, correct, or revise. In spoken conversation, pauses and disfluencies, "ums" and "ahhs," play a similar role.

[14] George Saunders, *A Swim in the Pond in the Rain* (New York: Bloomsbury, 2001).

[15] Daniel Adiwardana, Minh-Thang Luong, David R. So, et al., "Towards a Human-Like Open-Domain Chatbot," arXiv (2020), https://arxiv.org/abs/2001.09977.

[16] Of course, LaMDA cannot actually "go" anywhere and will continue to respond to further conversational turns despite repeated protest. Still, it can feel abusive to press on in these circumstances.

[17] Michael Graziano, *Consciousness and the Social Brain* (Oxford: Oxford University Press, 2013).

[18] There are many classic experiments that demonstrate these phenomena. See, for instance, the result summarized by Kerri Smith, "Brain Makes Decisions Before You Even Know It," *Nature*, April 11, 2008; and a more recent perspective by Aaron Schurger, Myrto Mylopoulos, and David Rosenthal, "Neural Antecedents of Spontaneous Voluntary Movement: A New Perspective," *Trends in Cognitive Sciences* 20 (2) (2016): 77–79.

[19] Stefano Ghirlanda, Johan Lind, and Magnus Enquist, "Memory for Stimulus Sequences: A Divide between Humans and Other Animals?" *Royal Society Open Science* 4 (6) (2017): 161011.

[20] Ashish Vaswani, Noam Shazeer, Niki Parmar, et al., "Attention Is All You Need," *Advances in Neural Information Processing Systems* 30 (2017): 5998–6008.

[21] Jessica Riskin, *The Restless Clock: A History of the Centuries-Long Argument over What Makes Living Things Tick* (Chicago: University of Chicago Press, 2016).

[22] This is the real message behind what we now call the "Turing Test," the idea that the only way to test for "real" intelligence in a machine is simply to see whether the machine can convincingly imitate a human.

Signs Taken for Wonders: AI, Art & the Matter of Race

Michele Elam

AI shares with earlier socially transformative technologies a reliance on limiting models of the "human" that embed racialized metrics for human achievement, expression, and progress. Many of these fundamental mindsets about what constitutes humanity have become institutionally codified, continuing to mushroom in design practices and research development of devices, applications, and platforms despite the best efforts of many well-intentioned technologists, scholars, policy-makers, and industries. This essay argues why and how AI needs to be much more deeply integrated with the humanities and arts in order to contribute to human flourishing, particularly with regard to social justice. Informed by decolonial, disability, and gender critical frameworks, some AI artist-technologists of color challenge commercial imperatives of "personalization" and "frictionlessness," representing race, ethnicity, and gender not as normative self-evident categories nor monetized data points, but as dynamic social processes always indexing political tensions and interests.

> As he grew accustomed to the great gallery of machines, he began to feel the forty-foot dynamos as a moral force, much as the early Christians felt the Cross. The planet itself felt less impressive, in its old-fashioned, deliberate, annual or daily revolution, than this huge wheel, revolving within arm's length at some vertiginous speed, and barely murmuring – scarcely – humming an audible warning to stand a hair's breadth further for respect of power, while it would not wake the baby lying close against its frame. Before the end, one began to pray to it; inherited instinct taught the natural expression of man before silent and infinite force. Among the thousand symbols of ultimate energy the dynamo was not so human as some, but it was the most expressive.
>
> —Henry Adams, "The Virgin and the Dynamo"[1]

In astonishment of the new technologies at the turn into the twentieth century, the renowned historian Henry Adams found the Gallery of the Electric Machines "physics stark mad in metaphysics" and wondered at their profound hold on the cultural imagination.[2] The dynamo that so moved and unset-

© 2022 by Michele Elam
Published under a Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) license
https://doi.org/10.1162/DAED_a_01910

tled Adams was a new generator of unprecedented scale, a machine responsible for powering the first electrified world's fair in 1893, a purportedly spectacular event presided over by President Glover Cleveland. Its power was invisible but the more potent for it: "No more relation could he discover between the steam and the electric current than between the cross and the cathedral. The forces were interchangeable if not reversible, but he could see only an absolute fiat in electricity as in faith." For Adams, the dynamo's effect in the world was akin to evidence of things unseen like the symbols of the Virgin or the cross, imperceptible but world-transforming currents with implications both worldly and spiritual.

I open with this discussion of the world's fair at the fin de siècle because Adams's dynamo is our GPT-3 (Generative Pre-trained Transformer 3), a language model that uses deep learning to produce text/speech/responses that can appear generated by a human. His exhilaration – hand-in-glove with his existential vertigo – and his internal conflict similarly speak to our contemporary aspirations for and anxieties about artificial intelligence. Adams understood that the turn to such formidable technology represented a thrilling but cataclysmic event, "his historical neck broken by the sudden irruption of forces entirely new." Although human grappling with exponential leaps in technology dates at least to the medieval period, this particular historical precedent of a transformational moment is singularly relevant for our contemporary moment: there's a direct line between Adams's concern with the hagiography of tech, the devaluation of the arts and humanities, and the comingling of scientific development with (racialized, ableist) narratives of progress to current debates about those nearly identical phenomena today. The consequences of those fundamental mindsets and practices, institutionally codified over time, continue to mushroom in devices, applications, platforms, design practices, and research development. Unacknowledged or misunderstood, they will continue to persist despite the best efforts of many well-intentioned technologists, scholars, policy-makers, and industries that still tend to frame and limit questions of fairness and bias in terms "safety," which can mute or obscure attention to issues of equity, justice, or power.[3]

Significantly, Adams's response to the dynamo is neither apocalyptic jeremiad nor in the genre of salvation: that is, his concerns fell beyond the pale of narratives of dystopia or deliverance. He was no technophobe; in fact, he deeply admired scientific advances of all kinds. Rather, his ambivalence has to do with the *inestimable psychological and spiritual sway* of machines so impressive that "the planet itself felt less impressive," even "old-fashioned."[4] That something man-made might seem so glorious as to overshadow creation, seemed so evocative of the infinite that people felt out of step with their own times. For Adams, those experiences signaled an epistemic break that rendered people especially receptive and open to change, but also vulnerable to idolizing false gods of a sort. He saw that the dynamo was quickly acquiring a kind of cult status, inviting supplication and reverence

Figure 1
Gallery of the Electric Machines, The Great Exposition, 1900 Paris World's Fair



Source : La Galerie des Machines Électriques at the Fifth Paris International Exposition of 1900. Image from Dynamo Exhibition Gallery of France, https ://www.ndl.go.jp/exposition/e/data/L/428l.html.

by its followers. The latest technology, as he personified it in his poem "Prayer to the Dynamo," was simultaneously a "Mysterious Power ! Gentle Friend ! Despotic Master ! Tireless Force !"[5] Adams experienced awe in the presence of the dynamo : "awe" as the eighteenth-century philosopher Edmund Burke meant the term, as being overcome by the terror and beauty of the sublime. And being tech awe-struck, he also instantly presaged many of his generation's – and I would argue, our generation's – genuflection before it.

As part of his concern that sophisticated technology inspires a kind of secular idolatry, Adams also noted its increasing dominance as the hallmark of human progress. In particular, he presciently anticipated that it might erode the power of

both religion and the arts as vehicles for and markers of humanity's higher strivings. Indeed, his experience at the Gallery taught him firsthand how fascination with such potent technology could eclipse appreciation of the arts: more specifically, of technological innovation replacing other modes of creative expression as the pinnacle of human achievement. Adams bemoaned the fact that his friend, Langley, who joined him at the exposition, "threw out of the field every exhibit that did not reveal a new application of force, and naturally, to begin with, the whole art exhibit." The progress of which technology increasingly claimed to be the yardstick extended beyond the valuation of art also extended to racial, ethnic, and gender scales. Most contemporary technological development, design, and impact continue to rely unquestioningly on enlightenment models of the "human," as well as the nearly unchanged and equally problematic metrics for human achievement, expression, and progress.

These are not rhetorical analogies; they are antecedences to AI, historical continuities that may appear obscured because the tech-ecosystem tends to eschew history altogether: discourses about AI always situate it as future-facing, prospective not retrospective. It is an idiom distinguished by incantations about growth, speed, and panoptic capture. The messy, recursive, complex narratives, events, and experiences that actually make up histories are reduced to static data points necessary in training sets for predictive algorithms. Adams's reaction offers an alternative framing of time in contrast to marketing imperatives that fetishize the next new thing, which by definition sheds its history.

This reframing is important to note because for all the contemporary talk of disruption as the vaulted and radical mode of innovation, current discourse still often presents so-called disruptive technologies as a step in an inexorable advance forward and upward. In that sense, tech disruption is in perfect keeping with the same teleological concept of momentum and progress that formed the foundational basis by which world's fairs ranked not only modes of human achievement but also degrees of "human." The exhibitions catalogued not just inventions but people, classifying people by emerging racialized typologies on a hierarchical scale of progress with the clear implication that some were more human than others.[6] This scale was made vivid and visceral: whether it was the tableaux vivant "ethnic villages" of the 1893 world's fair in Chicago's "White City" or the 1900 Paris showcase of African American achievement in the arts, humanities, and industry (images of "racial uplift" meant to counter stereotyping), both recognized how powerfully influential were representations of races' putative progress – or lack of it.

Carrying the international imprimatur of the fairs, the exhibitions were acts of racial formation, naturalizing rungs of humanness and, indeed, universalizing the imbrication of race and progress. Billed as a glimpse into the future, the fairs simultaneously defined what was *not* part of modernity: what or who was irrelevant, backward, regressive in relation. Technological progress, therefore, was not

simply represented *alongside* what (arts/humanities) or who (non-whites) were considered less progressive; progress was necessarily measured *against* both, indeed constituted by its difference and distance from both.

For critical theorist Homi Bhabha, such notions of progress, and the technology and symbol of it, are inextricably tied to the exercise of colonial and cultural power. His essay "Signs Taken for Wonders: Questions of Ambivalence and Authority Under a Tree outside Delhi, May 1817" critiques the "wondrous" presence of the book, itself a socially transformative technology, by beginning with the premise that innovation cannot be uncoupled from the prerogatives of those who have the power to shape realities with it:

> The discovery of the book is, at once, a moment of originality and authority, as well as a process of displacement, that paradoxically makes the presence of the book wondrous to the extent to which it is repeated, translated, misread, displaced. It is with the emblem of the English book – "signs taken as wonders" – as an insignia of colonial authority and an insignia of colonial desire and discipline that I begin this essay.[7]

Adams spoke of awe in the presence of the dynamo. Bhabha goes further in challenging such "signs taken as wonders," in questioning technologies so valorized that they engender awe, obeyance, and reverence as if such a response was natural, innocent of invested political and economic interests, free of market value systems.

> Like all tools, AI challenges the notion that the skull marks the border of the mind. . . . New tools breed new literacies, which can engender nascent forms of knowing, feeling and telling.
>
> —Vanessa Chang, "Prosthetic Memories, Writing Machines"[8]

Art sits at the intersection of technology, representation, and influence. Literature, film, music, media, and visual and graphic arts are all crucial incubators for how publics perceive tech. Storytelling impacts, implicitly or explicitly, everything from product design to public policy. Many of these narratives bear traces of literature's earliest engagement with technology, at least since medieval times, and others – either engaged with AI or AI-enabled – are also offering new plotlines, tropes, identity formations, historiographies, and speculative futurities. Moreover, because cultural storytelling helps shape the civic imagination, it can, in turn, animate political engagement and cultural change.[9]

Indeed, the arts are specially poised to examine issues in technological spaces (from industry to STEM education) of equity, diversity, social justice, and power more capaciously and cogently than the sometimes reductive industry-speak of inclusion, fairness, or safety (usually simply meaning minimization of harm

or death – a low bar indeed). Even before GPT-3, powerful natural language processing was enabling explorations in AI-assisted poetry, AI-generated film-scripts, AI-informed musicals, AI-advised symphonies, AI-curated art histories, and AI-augmented music.[10] Many are proposing new nomenclature for hybrid genres of art, design, and tech, and fresh subfields are blooming in both academe and entertainment.[11] And during the COVID-19 pandemic and intensified movements for social justice, there has been a plethora of virtual exhibitions and articles about the hot debates over the status, meaning, and valuation of AI-generated or -augmented art.[12]

Amidst this explosion of artistic engagement with AI, social and political AI scholars Kate Crawford and Luke Stark, in "The Work of Art in the Age of Artificial Intelligence: What Artists Can Teach Us about the Ethic of Data Practice," offer a not uncommon perspective on the need for interdisciplinary collaboration: "Rather than being sidelined in the debates about ethics in artificial intelligence and data practices more broadly, artists should be centered as practitioners who are already seeking to make public the political and cultural tensions in using data platforms to reflect on our social world."[13] However, they also close the article by recommending that arts practitioners and scholars would do well with more technical education and that without it, their engagements and critiques will have lesser insight into and standing regarding the ethics of data practice: "One barrier to a shared and nuanced understanding of the ethical issues raised by digital art practices is a lack of literacy regarding the technologies themselves. . . . Until art critics engage more deeply with the technical frameworks of data art, their ability to analyze and assess the merits of these works – and their attendant ethical dilemmas – may be limited." They suggest: "a close relationship to computer science seemed to offer some artists a clearer lens through which to consider the ethics of their work."[14]

Certainly, continuing education is usually all to the good. But I would welcome the equivalent suggestion that those in data science, computer science, engineering, and technology, in turn, should continue to educate *themselves* about aesthetics and arts practices – including at least a passing familiarity with feminist, queer, decolonial, disability, and race studies approaches to AI often central to those practices – to better understand ethical debates in their respective fields.[15] Without that balance, the suggestion that artists and nontechnical laypeople are the ones who primarily need education, that they require technical training and credentialing in order to have a valid(ated) understanding of and legitimate say in the political, ethical, social, and economic discussions about AI, is a kind of subtle gatekeeping that is one of the many often unacknowledged barriers to cross-disciplinary communication and collaboration. Given the differential status of the arts in relation to technology today, it is usually taken for granted that artists (not technologists, who presumably are doing more important and time-consuming

work in and for the world) have the leisure and means not only to gain additional training in other fields but also to do the hard translational work necessary to integrate those other often very different disciplinary practices, vocabularies, and mindsets to their own creative work. That skewed status impacts who gains the funding, influence, and means to shape the world.

Instead of asking artists to adapt to the world models and pedagogies informing technological training – which, as with any education, is not simply the neutral acquisition of skills but an inculcation to very particular ways of thinking and doing – industry might do well to adapt to the broader vernacular cultural practices and techne of marginalized Black, Latinx, and Indigenous communities. Doing so might shift conversation in the tech industry from simply mitigating harm or liability from the differentially negative impact of technologies on these communities. Rather, it would require a mindset in which they are recognized as equal partners, cultural producers of knowledge(s), as the longtime makers, not just the recipients and consumers, of technologies.[16] In fact, artist-technologist Amelia Bearskin-Winger, who is Haudenosaunee (Iroquois) of the Seneca-Cayuga Nation of Oklahoma, Deer Clan, makes a case that many of these vernacular, often generational, practices and values are what she calls "antecedent technologies," motivated by an ethic that any innovation should honor its debt to those seven generations prior and pay it forward seven generations.[17]

In this way, many contemporary artist-technologists engage issues including, but also going beyond, ethics to explore higher-order questions about creativity and humanity. Some offer non-Western or Indigenous epistemologies, cosmologies, and theologies that insist on rethinking commonly accepted paradigms about what it means to be human and what ways of doing business emerge from that. Perhaps most profoundly, then, the arts can offer different, capacious ways of knowing, seeing, and experiencing worlds that nourish well-being in the now and for the future. It is a reminder of and invitation to world models and frameworks alternative to what can seem at times to be dominating or totalizing technological visions. In fact, one of the most oft-cited criticisms of AI discourse, design, and application concerns its univision, its implied omniscience, what scholar Alison Adams calls "the view from nowhere." It is challenged by art that offers simultaneous, multiple, specifically situated, and sometimes competing points of view and angles of vision that enlarge the aperture of understanding.[18]

For instance, informed by disability culture, AI-augmented art has drawn on GANs (generative adversarial networks) to envision non-normative, including neurodivergent, subjects that challenge taken-for-granted understandings of human experience and capability. The presumption of a universal standard or normative model, against which "deviance" or "deviation" is measured, is nearly always implied to be white, cis-gendered, middle-classed, and physically and cognitively abled. That fiction of the universal subject – of what disability scholar

and activist Rosemarie Garland-Thomson terms the "normate" – has historically shaped everything from medical practice and civil rights laws to built environments and educational institutions. It also often continues to inform technologies' development and perceived market viability and use-value. Representations of "human-centered" technology that include those with mental or physical disabilities often call for a divestment from these usual ways of thinking and creating. Such a direct critique is posed in art exhibitions such as *Recoding CripTech*. As the curatorial statement puts it, the installations reimagine "enshrined notions of what a body can be or do through creative technologies, and how it can move, look or communicate. Working with a broad understanding of technology . . . this multidisciplinary community art exhibition explores how disability – and artists who identify as such – can redefine design, aesthetics and the relationship between user and interface." Works included in *Recoding CripTech* that employ artificial intelligence, such as M Eifler's "Prosthetic Memory" and "Masking Machine," suggest a provocative reframing of "optimization" or "functionality" in technologies that propose to augment the human experience.[19]

> Race – racism – is a device. No More. No less. It explains nothing at all. . . . It is simply a means. An invention to justify the rule of some men over others. [But] it also has consequences; once invented it takes on a life, a reality of its own. . . . And it is pointless to pretend that it doesn't exist – merely because it is a lie!
>
> —Tshembe in *Les Blancs* (1965) by Lorraine Hansberry

Rashaad Newsome's installation *Being* represents another artistic provocation that reframes both the form and content of traditional technological historiographies often told from that "view from nowhere." Newsome, a multimedia artist and activist, makes visible the erased contributions to technology and art by people of African descent. Newsome terms the interactive social humanoid *Being* 2.0 an "AI griot," a storyteller. But unlike most social robots commanded to speak, *Being* is intentionally "uppity": wayward, noncompliant, disobedient, with expressive gestures drawn Black Queer vogue dance repertoire meant as gestures of decolonial resistance to the labor and service that social robots are expected to perform. It upends the historical association of robots and slaves (in the etymology of the Czech word, "robot" translates to "slave") in movement, affect, function, and speech. Taking aim at the limited training data sets used in natural language processing, Newsome draws on broader archives that include African American vernacular symbolic systems.[20] And since language carries cultural knowledge, *Being*'s speech expands not just vocabularies but reimagines how the standardized expressions of emotion and behavior often deployed in AI are racially and culturally encoded.[21] In fact, *Being* is an attempt to redress the historical

Figure 2
Rashaad Newsome's *Being* 2.0



Being © Rashaad Newsome Studio.

violence of antiquated notions about race, the more disturbing because the representations of race, reduced to seemingly self-evident graduations of color and physiognomy, are being actively resurrected in AI development and application.

Race is always a negotiation of social ascription and personal affirmation, a process of what sociologists Michael Omi and Howard Winant term "racial formation." Omi and Winant refer to racial formation as a way of historicizing the practices and circumstances that generate and renew racial categories and racializing structures:

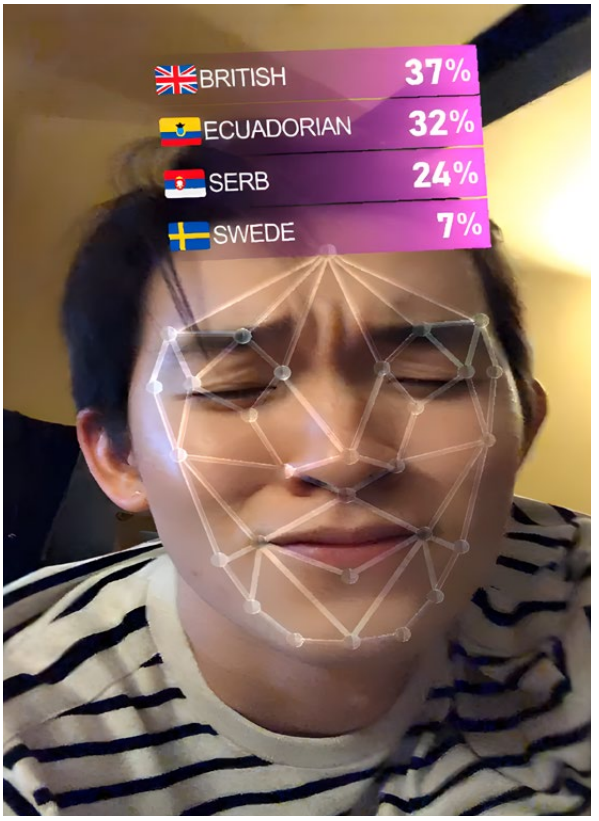
> We define *racial formation* as the sociohistorical process by which racial categories are created, inhabited, transformed, and destroyed. . . . Racial formation is a process of historically situated *projects* in which human bodies and social structures are represented and organized. Next we link racial formation to the evolution of hegemony, the way in which society is organized and ruled. . . . From a racial formation perspective, race is a matter of both social structure and cultural representation.[22]

The expression "racial formation" is therefore a reminder that race is not *a priori*. It is a reminder to analyze the structural and representational – not just linguistic – contexts in which race becomes salient: the cultural staging, political investments, institutional systems, and social witnessing that grant meanings and values to categories. A full accounting of race therefore involves asking in whose interest is it that a person or people are racialized in any given moment in time and space? What function does it enable or disable? In short, what does it *do*? As Toni Morrison asks, the question should not be simply "is it racist?" but rather what does race – its presence or its conspicuous absence – make possible or eclipse?[23]

Overlooked, for instance, in many debates over racial bias, surveillance, and privacy in facial recognition technology is the practice of coding "race" or "ethnicity" as fixed, static programmable variables, something writ on the face or otherwise available as physically intelligible – an outdated approach to race that harkens back to nineteenth-century phrenology and other pseudoscience mappings of racial traits. Moreover, that practice renders opaque how categories are never merely descriptive, disinterested renderings of facts or things even though they cannot be purged of the value systems that animate their creation and make them intelligible for technological use – at least as currently developed – in the first place. Additionally, the claim to a universal objectivity is one of the "epistemic forgeries," according to Yarden Katz, who describes it as one of the "fictions about knowledge and human thoughts that help AI function as a technology of power" because it enables "AI practitioners' presumption that their systems represent a universal 'intelligence' unmarked by social context and politics."[24] That drive for comprehensive typing and classification, for a universal compendium, cannot easily accommodate race other than a technical problem in mapping variation of types.[25]

To illustrate why AI representations are so problematic, let me take a seemingly innocuous example in the new algorithmic application "Ethnicity Estimate," part of the Gradient app, which purports to diagnose percentages of one's ethnic heritage based on facial recognition technology (FRT). Such an app is significant precisely because popular data-scraping applications are so often pitched as convenient business solutions or benign creative entertainment, bypassing scrutiny because they seem so harmless, unworthy of research analysis or quantitative study. Critically examining on such issues would be a direct impediment to

Figure 3
Ethnicity Estimate and Facial Recognition Technology

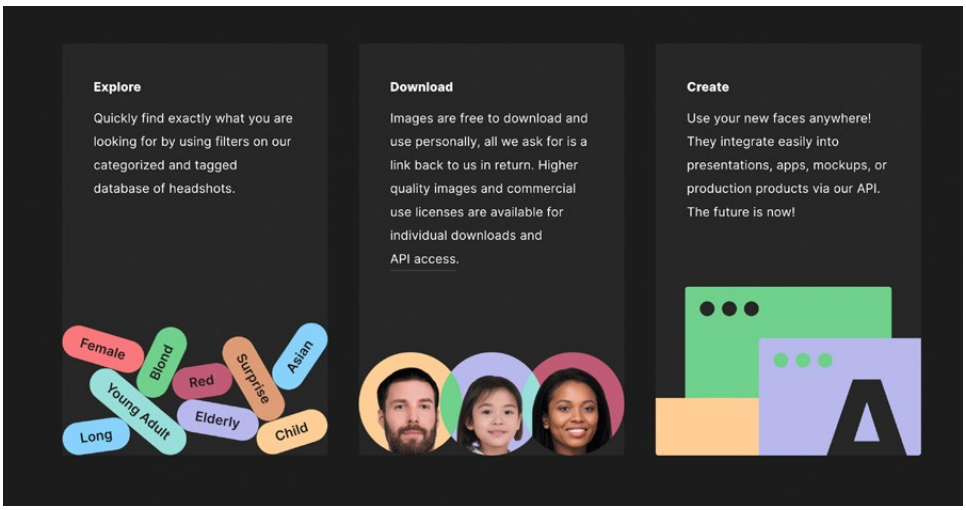


Screenshot of the Ethnicity Estimate tool using FRT on one of my students, Edric Zeng, who is Korean and Chinese. Note his incredulous expression upon seeing its conclusion : 37 percent British ; 32 percent Ecuadorian ; 24 percent Serb ; 7 percent Swede. Image courtesy of Edric Zeng.

a seamless user experience with the product, thus designers and users are actively disincentivized from doing so. Like many such applications, Ethnicity Estimate problematically uses nationality as a proxy for ethnicity and reduces population demographics to blood quantum.

Or consider Generated Photos : an AI-constructed image bank of "worry-free" and "infinitely diverse" facial portraits of people who do not exist in the flesh, which marketers, companies, and individuals can use "for any purpose without worrying about copyrights, distribution rights, infringement claims or royalties."[26] In creating these virtual "new people," the service offers a workaround for privacy concerns. Generated Photos bills itself as the future of intelligence, yet it reinscribes

Figure 4
Generated Photos: "Use Your New Faces Anywhere!"



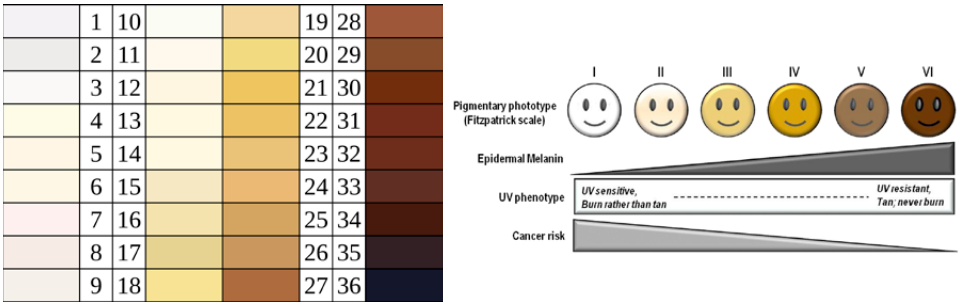
Source: Screenshot of promotional materials on https://generated.photos/.

the most reductive characterizations of race: among other parameters users can define when creating the portraits, such as age, hair length, eye color, and emotion through facial expression, the racial option has a dropdown of the generic homogenizing categories Asian, African American, Black, Latino, European/white.

Skin color options are similarly presented as self-evident and unproblematic givens, a data set based on an off-the-shelf color chart. There is a long racializing history of such charts, from the von Luschan chromatic scale, used throughout the first half of the twentieth century to establish racial classifications, to the Fitzpatrick scale, still common in dermatologists' offices today, which classifies skin types by color, symbolized by six smiling emoji modifiers. Although the latter makes no explicit claim about races, the emojis clearly evoke the visuals well as the language of race with the euphemism of "pigmentary phototype."

All these types are readily serviceable as discrete data points, which makes them an easy go-to in algorithmic training, but the practice completely elides the fact that designations of "dark" or "light" are charged cultural and contextual interpretations that are always negotiated in context and *in situ*.[27] The relevance and meaning of race emerge through social and cultural relations, not light frequencies. Fran Ross's brilliant, satirical novel *Oreo* (1974) offers a wry send-up of attempts to apply color charts to social identities, shown as Figure 6.[28]

Figure 5
The von Luschan Chromatic Scale *(left)* and the Fitzpatrick Scale *(right)*



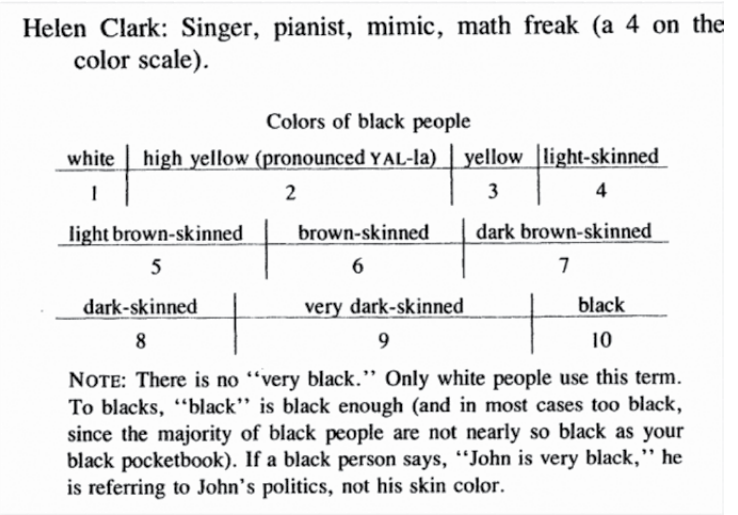
The reproduction of the von Luschan chromatic scale, based on the chart first printed in *Völker, Rassen, Sprachen* (1927), is by Wikimedia users Dark Tichondrias and Cburnett. Printed under the Creative Commons Attribution-ShareAlike 3.0 Unported license. The Fitzpatrick scale is from John D'Orazio, Stuart Jarrett, Alexandra Amaro-Ortiz, and Timothy Scott, "UV Radiation and the Skin," *International Journal of Molecular Sciences* 14 (6) (2013). Reprinted under the Creative Commons Attribution 3.0 Unported license.

Table 1
Fitzpatrick Type and von Luschan Scale

Fitzpatrick Type	Von Luschan Scale	Also Called
I	0–6	Very light or white, "Celtic" type
II	7–13	Light or light-skinned European
III	14–20	Light intermediate, or dark-skinned European
IV	21–27	Dark intermediate or "olive skin"
V	28–34	Dark or "brown" type
VI	35–36	Very dark or "black" type

Source : Nina G. Jablonski, "Skin Coloration" in *Human Evolutionary Biology*, ed. Michael P. Muehlenbein (Cambridge : Cambridge University Press, 2010), 177.

Figure 6
Fran Ross's *Oreo* Color Scale



Helen Clark: Singer, pianist, mimic, math freak (a 4 on the color scale).

Colors of black people

white	high yellow (pronounced YAL-la)	yellow	light-skinned
1	2	3	4

light brown-skinned	brown-skinned	dark brown-skinned
5	6	7

dark-skinned	very dark-skinned	black
8	9	10

NOTE: There is no "very black." Only white people use this term. To blacks, "black" is black enough (and in most cases too black, since the majority of black people are not nearly so black as your black pocketbook). If a black person says, "John is very black," he is referring to John's politics, not his skin color.

Source: Fran Ross, *Oreo* (Boston: Northeastern University Press, 1974).

Although new AI technologies show promise in diagnosing medical conditions of the skin, thinking of racial identification primarily in terms of chromatic scales or dermatoscopic data deflects attention, to put it generously, from the long history of the damaging associations of skin color and race that gave rise to early technologies like this in the first place, whether it was the "science" of phrenology, IQ tests, or fingerprinting, and with implications, more recently, for the use of biometrics.[29] At a minimum, it ignores the imbrication of "race" in pigmentocracies and colorism, the historical privileging of light skin, and the various rationales for identifying what counts as "light-skinned." Colorism, a legacy of colonialism, continues to persist in contemporary hierarchies of value and social status, including aesthetics (who or what is ranked beautiful, according to white, Western standards), moral worth (the religious iconography of "dark" with evil and "light" with holy continues to saturate languages), social relations (for instance, the "paper bag test" of the twentieth century was used as a form of class gatekeeping in some African American social institutions),[30] and the justice system (since social scientists have documented the perceptual equation of "blackness" with crime, and thus those perceived as having darker skin as *a priori* criminally suspect).[31]

Why does this matter? Because it suggests that the challenges in representing race in AI are not something technological advances in any near or far future

could solve. Rather, they signal cultural and political, not technical, problems to address. The issue, after all, is not merely a question of bias (implicit or otherwise), nor of inaccuracy (which might lead some to think the answer is simply the generation of more granular categories), nor of racial misrecognition (which some might hear as simply a call for ever more sophisticated FRT), nor even of ending all uses of racial categorization.[32] It matters because algorithms trained on data sets of racial types reinforce color lines, literally and figuratively remanding people back in their "place." By contrast, as I have suggested, the increasingly influential rise of AI artist-technologists, especially those of color, are among those most dynamically questioning and reimagining the commercial imperatives of "personalization" and "frictionlessness." Productively refusing colorblindeness, they represent race, ethnicity, and gender not as normative, self-evident categories nor monetizable data points, but as the dynamic social processes – always indexing political tensions and interests – which they are. In doing so, they make possible the chance to truly create technologies for social good and well-being.

> Something has happened. Something very big indeed, yet something that we have still not integrated fully and comfortably into the broader fabric of our lives, including the dimensions – humanistic, aesthetic, ethical and theological – that science cannot resolve, but that science has also (and without contradiction) intimately contacted in every corner of its discourse and being.
>
> —Stephen Jay Gould, *The Hedgehog, the Fox, and the Magister's Pox* (2003)[33]

I cite what may seem minor examples of cultural ephemera because, counterintuitively, they hint at the grander challenges of AI. They are a thread revealing the pattern of "something very big indeed," as historian of science Stephen Jay Gould put it. Certainly there are ethical, economic, medical, educational, and legal challenges facing the future of AI. But the grandest technological challenge may in fact be *cultural*: the way AI is shaping the human experience. Through that lens, the question becomes not one of automation versus augmentation, in which "augmenting" refers to economic productivity, but rather to creativity. That is, how can AI best augment the arts and humanities and thus be in service to the fullness of human expression and experience?

This essay opened with Henry Adams's moment of contact with the Dynamo's "silent and infinite force," as he put it, which productively *de*naturalizes the world as he knows it, suspends the usual epistemological scripts about the known world and one's place in it. It is a sentiment echoed almost verbatim two hundred years later by Gould, witnessing another profound technological and cultural upending. Writing at the turn into our own century, Gould, like Adams, cannot fully articulate the revelation except to say poignantly that "something has happened,"

that every dimension of "the broader fabric of our lives" is intimately touched by a technology whose profound effect cannot be "solved" by it. That liminal moment for Adams, for Gould, and for us makes space for imagining other possibilities for human creativity, aesthetic possibilities that rub against the grain and momentum of current technological visions, in order to better realize the "magisteria of our full being."[34]

ABOUT THE AUTHOR

Michele Elam is the William Robertson Coe Professor of Humanities in the Department of English, Faculty Associate Director of the Institute for Human-Centered Artificial Intelligence, the Bass University Fellow in Undergraduate Education, and a Race & Technology Affiliate at the Center for Comparative Studies in Race and Ethnicity at Stanford University. She is the author of *The Souls of Mixed Folk: Race, Politics, and Aesthetics in the New Millennium* (2011) and *Race, Work, and Desire in American Literature, 1860 – 1930* (2003) and editor of *The Cambridge Companion to James Baldwin* (2015).

ENDNOTES

[1] Henry Adams, "The Virgin and the Dynamo," in *The Education of Henry Adams* (Boston : self published, 1907). All of Henry Adams's quotes are taken from the unpaginated PDF hosted by the University of Chicago at http://geosci.uchicago.edu/~moyer/GEOS24705/2009/HenryAdams.pdf.

[2] Unlike in *The Wonderful Wizard of Oz*, L. Frank Baum's children's book published the same year as the Paris Exhibition, for Adams, there is no reveal of the "man behind the curtain," no Oz orchestrating a show. His interest is not in the technologists but in what ontological truths their creations tap.

[3] While there is no clear ethical or legal consensus on what constitutes "fairness," there are critiques of fairness models that assume an equal playing field thwarting access to opportunities, that presume equal discrimination equals fairness, or that understand fairness in the narrowest sense of preventing harm (the critique of situating "fairness" under concerns of "safety"). For a review of some of the debates about fairness, see Michele Elam and Rob Reich, "Stanford HAI Artificial Intelligence Bill of Rights" (Stanford, Calif. : Stanford Institute for Human-Centered Artificial Intelligence, 2022), https://hai.stanford.edu/sites/default/files/2022-01/Stanford%20HAI%20Artificial%20Intelligence%20Bill%20of%20Rights_0.pdf.

[4] Adams, "The Virgin and the Dynamo."

[5] The "Prayer to the Dynamo" is a section within a longer poem by Henry Adams, *Prayer to the Virgin of Chartres* (1900), accessed through Teaching American History, https://teachingamericanhistory.org/library/document/prayer-to-the-virgin-of-chartres/.

[6] On the question of humanness, and for an analysis of how technologies determine full humans, not-quite-humans, and nonhumans, see Alexander G. Weheliye's excellent *Habeas Viscus: Racializing Assemblages, Biopolitics, and Black Feminist Theories of the Human* (Durham, N.C.: Duke University Press, 2014). See also Sylvia Wynter's foundational essay, "Unsettling the Coloniality of Being/Power/Truth/Freedom: Towards the Human, After Man, Its Overrepresentation–An Argument," *CR: The New Centennial Review* 3 (3) (2003): 257–337. Wynter critiques the overrepresentation of man (as white, Western) as the only imaginable mode of humanness, overwriting other ontologies, epistemologies, and imaginaries. See also Katherine McKittrick, ed., *Sylvia Wynter: On Being Human as Praxis* (Durham, N.C.: Duke University Press, 2015). A major influence on this essay, Wynter's pioneering and prolific work draws on arts, humanities, natural sciences and neuroscience, philosophy, literary theory, and critical race theory.

[7] Homi Bhabha, "Signs Taken for Wonders: Questions of Ambivalence and Authority Under a Tree outside Delhi, May 1817," *Critical Inquiry* 12 (1) (1985): 144.

[8] Vanessa Chang, "Prosthetic Memories, Writing Machines," *Noëma*, December 3, 2020, https://www.noemamag.com/prosthetic-memories-writing-machines/.

[9] For instance, see the comment made by the then head of Instagram, Adam Mosseri, that his recent policy eliminating public "likes"–because of his concern "about the unintended consequences of Instagram as an approval arbiter" as he put it–was partly informed by an episode of the science fiction anthology television series *Black Mirror*. Amy Chozick, "This Is the Guy Who's Taking Away the Likes," *The New York Times*, January 17, 2020, https://www.nytimes.com/2020/01/17/business/instagram-likes.html.

[10] For example, see Google's "Poem Portraits by Es Devlin," https://artsexperiments.withgoogle.com/poemportraits?_ga=2.33161846.992826029.1556786810-799000725.1554196893. A 2016 short-film script entitled *Sunspring* was made by an AI bot. The main character in the filmed version was played by Thomas Middleditch, the same actor who plays the lead, Richard Hendriks, in the TV series *Silicon Valley* 😊. See also AI-generated musicals and symphonies, as in Maura Barrett and Jacob Ward, "AI Can Now Compose Pop Music and Even Symphonies. Here's How Composers Are Joining In," MACH by NBC News, May 29, 2019, https://www.nbcnews.com/mach/science/ai-can-now-compose-pop-music-even-symphonies-here-s-ncna1010931.

[11] Artist-technologists working at these intersections include Amelia Bearskin-Winger, Legacy Russell, Stephanie Dinkins, Ian Chang, Rashaad Newsome, Jacolby Satterwhite, Joy Buolamwini, Martine Syms, and others–not to mention those writers in the long literary history of speculative fiction.

[12] See, for instance, "Art at a Time Like This," https://artintimeslikethis.com/. I take up in detail the debates over the value and valuation of the arts and humanities in the age of artificial intelligence in Michele Elam, "GPT-3 in 'Still I Rise!': Why AI Needs Humanists," *American Literature* (forthcoming spring 2022).

[13] Luke Stark and Kate Crawford, "The Work of Art in the Age of Artificial Intelligence: What Artists Can Teach Us about the Ethics of Data Practice," *Surveillance & Society* 17 (3/4) (2019): 452.

[14] Ibid., 451–452.

[15] There are many academic and activist resources and collectives working in these areas, including Latinx in AI, Black in AI, Queer AI, Indigenous AI, and Accessible AI, to name but a few.

[16] See, for instance, Ruha Benjamin, "Introduction: Discriminatory Design, Liberating Imagination," in *Captivating Technology: Race, Carceral Technoscience, and Liberatory Imagination in Everyday Life*, ed. Ruha Benjamin (Durham, N.C.: Duke University Press, 2019), 1–25; and Nettrice R. Gaskins, "Creativity and Innovation across the African Diaspora and Global South," in *Captivating Technology*, 252–253.

[17] Amelia Winger-Bearskin explains the concept in "Antecedent Technology: Don't Colonize Our Future," *Immerse News*, November 21, 2019, https://immerse.news/antecedent-technology-b3a89956299d; and Amelia Winger-Bearskin, "Before Everyone Was Talking about Decentralization, Decentralization Was Talking to Everyone," *Immerse News*, July 2, 2018, https://immerse.news/decentralized-storytelling-d8450490b3ee.

[18] Alison Adams quoted in Yarden Katz, *Artificial Whiteness: Politics and Ideology in Artificial Intelligence* (New York: Columbia University Press, 2020), specifically in Katz's discussion of AI notions of the self: "Practitioners in the 1970s, for instance, offered visions of the self as a symbolic processing machine. . . . In the late 1980s and early 1990s, by contrast, the prevailing 'self' started looking more like a statistical inference engine driven by sensory data. But these classifications mask more fundamental epistemic commitments. Alison Adams has argued that AI practitioners across the board have aspired to a 'view from nowhere'–to build systems that learn, reason, and act in a manner freed from social context. The view from nowhere turned out to be a view from a rather specific, white, and privileged space." Ibid., 6. See Alison Adams, *Artificial Knowing: Gender and the Thinking Machine* (Abingdon-on-Thames, United Kingdom: Routledge, 1998).

[19] See Rosemarie Garland-Thomson, *Extraordinary Bodies: Representing Physical Disability in American Culture and Literature* (New York: Columbia University Press, 1996). The curatorial statement in *Recoding CripTech* explains the terminology: "the term 'crip' reclaims the word for disability culture and recognizes disability as a political and cultural identity"; see *Recoding CripTech*, https://www.recodingcriptech.com/. See also M Eifler's *Prosthetic Memory* (2020), https://www.recodingcriptech.com/prosthetic-memory; and *Masking Machine* (2018), https://www.recodingcriptech.com/masking-machine. Many thanks to Lindsey Felt and Vanessa Chang, who produced and curated *Recoding CripTech*, and especially for Lindsey's suggestions for this section of the essay.

[20] Su Lin Blodge and Brendan O'Connor, "A Racial Disparity in Natural Language Processing: A Case Study in Social Media African American English," arXiv (2017), https://arxiv.org/pdf/1707.00061.pdf.

[21] See Neda Atanasoski and Kalindi Vora, "The Surrogate Human Affect: The Racial Programming of Robot Emotion," in *Surrogate Humanity: Race, Robots and the Politics of Technological Futures* (Durham, N.C.: Duke University Press, 2019), 108–133.

[22] Michael Omi and Howard Winant, *Racial Formation in the United States*, 2nd ed. (New York: Routledge, 1994), 55–56.

[23] Toni Morrison, *Playing the Dark: Whiteness and the Literary Imagination* (Cambridge, Mass.: Harvard University Press, 1992).

[24] Katz, *Artificial Whiteness: Politics and Ideology in Artificial Intelligence*. The claim to universality, according to Yatz, is the first forgery: "The second is that AI systems have matched or exceeded the capabilities of human thought . . . [drawing] on deep-seated notions in Western culture about hierarchies of intelligence. The third epistemic forgery suggests that these computational systems arrive at truth, or 'knowledge,' 'on their own,' AI practitioners being merely the ones who set off the necessary conditions for computational processes to properly unfold." Ibid., 94–95.

[25] Directly related to the issue of racial classification is the issue of racial and gender "ambiguity," often (mis)understood in AI as simply a technical issue of documenting (or not) variance and managing uncertainty. For extended discussion of the complicated technical and political challenges posed by social identities transgressing racial or gender boundaries in AI, see Michele Elam, "Recoding Gender-Binaries and Color-Lines : AI and the Erasure of Ambiguity" in *Feminist AI* (forthcoming 2022).

[26] Generated Photos, https://generated.photos/ (accessed December 17, 2021).

[27] By arguing for understanding words "in context," I mean in the social science sense of meaning emerging through performative interaction, not meaning garnered by sentence-level "context," as is commonly understood in natural language processing : that is, of other words that appear near each other in a sentence. Thanks to my Stanford colleagues Chris Manning and Surya Ganguli for our conversations about natural language processing.

[28] Fran Ross, *Oreo* (Boston : Northeastern University Press, 2000), 5. First published in 1974 by Greyfalcon House, Inc.

[29] This reduction of race to data points for AI has many problematic medical implications, even in work aiming to mitigate bias. Using self-reported racial data along with other medical information to create a "phenotype," Juan Banda acknowledges that racial porosity (people identifying with multiple races, differently over time, or differently in various contexts) as well as durational changes in aging mean up to 20 percent of data are thrown out in assessments. Moreover, as is long documented, race is often proxy for all sorts of other factors and inequities, so whether the data collected are self-reported or social ascription, it is still a highly problematic category. Juan Banda, "Phenotyping Algorithms Fair for Underrepresented Minorities within Older Adults," Stanford HAI Weekly Seminar with Juan Banda, February 2, 2022, https://philosophy.stanford.edu/events/hai-weekly-seminar-juan-banda-0.

[30] The "paper bag test," and color discrimination broadly, is a complicated form of internalized racism dating to slavery. Early scholarship suggested lighter-skinned slaves were supposedly treated better, although that view has been challenged, since being "in the Big House" instead of the field often meant greater exposure to sexual abuse. Moreover, there are many documented accounts of mixed race children–the issue of white masters and slaves–being sold away, often at the insistence of the white mistress of the house, since they stood as corporeal testimony to miscegenation and adultery–in short, to the sins of the father.

[31] See Jennifer Eberhardt, *Biased : Uncovering the Hidden Prejudice that Shapes What We See, Think, and Do* (London : Penguin, 2020). Eberhardt documents how police are more likely to think a suspect is holding a gun if they are perceived as "dark," and that juries are more likely to negatively regard, and thus weigh judgment against, a defendant if they are considered "dark," among the many other consequences.

[32] For instance, racial classification is often employed in the United States (though not in France and other countries) for strategic purposes generally understood as for the larger social good : for instance, by the Office of Management and Budget for census purposes, which are tied to the distribution of resources and the tracking of civil rights violations, or by Federal Statistical Research Data Centers that enable the study of health and economic trends across demographics. In such cases, data collection of those gross categories is undertaken with the understanding that these are efforts to capture not inherent traits or types, but rather broad trends in the complex of race, gender, and

socioeconomics, among other variables. More recently, efforts to collect immigration and citizen status, neither following the prior intent nor practice of the Census, has made vivid the potential for misuses of this national mode of data collection.

33 Stephen Jay Gould, *The Hedgehog, the Fox, and the Magister's Pox*: *Minding the Gap Between Science and the Humanities* (Cambridge, Mass.: Belknap Press, 2011), 15.

34 I am following Stephen Jay Gould's critiques in *The Hedgehog, the Fox, and the Magister's Pox* of E. O. Wilson and C. P. Snow's notion of two cultures, described in Snow's *The Two Cultures and the Scientific Revolution* as a split between science and humanities. Gould found that duality artificial, simplistic, and ahistorical, offering instead more accurate accounts of four historical periods exploring continuities of creative thinking across the sciences and humanities.

Toward a Theory of Justice for Artificial Intelligence

Iason Gabriel

This essay explores the relationship between artificial intelligence and principles of distributive justice. Drawing upon the political philosophy of John Rawls, it holds that the basic structure of society should be understood as a composite of sociotechnical systems, and that the operation of these systems is increasingly shaped and influenced by AI. Consequently, egalitarian norms of justice apply to the technology when it is deployed in these contexts. These norms entail that the relevant AI systems must meet a certain standard of public justification, support citizens' rights, and promote substantively fair outcomes, something that requires particular attention to the impact they have on the worst-off members of society.

Calls for justice in the context of artificial intelligence sound increasingly loud. Indeed, communications scholar Matthew Le Bui and gender studies scholar Safiya Umoja Noble have argued that we are missing a moral framework of justice altogether when it comes to evaluating the practices that constitute artificial intelligence.[1] The demand for justice represents both a need felt among those impacted by AI systems and a source of important philosophical insight. Among other things, it reframes much of the discussion around "AI ethics" by drawing attention to the fact that the moral properties of algorithms are not internal to the models themselves but rather a product of the social systems within which they are deployed. At the same time, those who want to evaluate emergent practices through the lens of justice rapidly encounter an obstacle: namely, that political theory – which is the body of thought we might hope to rely on to address these questions – has not adequately addressed technology in general, struggling to navigate a path between relative neglect and determinism. As a consequence, it is not necessarily well-equipped to speak to the role of technology in public life, let alone say something meaningful about justice and AI systems.

Taking these points in turn, much of contemporary political philosophy brackets out technological considerations, treating them as exogenous to the fundamental questions of political life. This view is found in the work of philosopher John Rawls, whose seminal work *A Theory of Justice* mentions *technology* on just three occasions. Moreover, although his account of justice appears to be for a society that

© 2022 by the American Academy of Arts & Sciences
Published under a Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) license
https://doi.org/10.1162/DAED_a_01911

has a specific sociotechnical character (that is, one with a functioning legal system, economic division of labor, capacity for taxation, and so on), knowledge about the level of technology that a society has achieved is excluded from the original position when selecting principles of justice. It is only when making a final assessment of what justice requires in specific contexts that we need to "take into account economic efficiency and the requirements of organization and technology."[2]

By contrast, technology plays a central role in Marxist thought. However, the account provided leaves little room for human choices or moral direction. Indeed, while the character of Marx's historical materialism remains subject to deep exegetical disagreement, one prominent interpretation holds that, for any given moment, the development of productive forces (that is, technology and labor) explains the nature of the mode of production (understood as the prevailing economic relations), which then shapes society's ideological superstructure, including its laws and system of beliefs.[3] Understood in this way, the development of technology still functions primarily as an exogenous force. Moreover, if prevalent moral norms are largely explained by material circumstances (and potentially nothing more than a "bourgeois ideology" in a late capitalist society), then they appear deeply, and perhaps terminally, compromised as a vantage point from which to make an independent moral evaluation.

Taken together, these accounts matter because they suggest that calls for justice in the context of AI are essentially misplaced. Understood primarily as a new technology, AI either falls outside the scope of justice or is part of a dynamic that prefigures robust moral evaluation. In this essay, I defend a different approach, one that makes claims about AI, justice, and injustice entirely appropriate.[4] This approach begins by noting that the interaction between humans and technology is a two-way process. On the one hand, we are profoundly affected by the technologies we adopt. In modern societies, technology helps to facilitate control from a single center, maintain larger organizational units, promote economic specialization, determine the meaning of authority and expertise, and shape the goals, aspirations, and self-understanding of citizens. On the other hand, we are not only acted upon by technologies, but we also create them through a process of design, experimentation, development, iteration, and adoption. Clearly, the power to shape and influence the path of technological change is not distributed evenly across society.[5] Nonetheless, choices about the content and character of new technologies are being made.

Taken together, what emerges therefore is a class of profound societal effects induced by technological change alongside a set of technological choices that shape the path of innovation via the decisions of individual technologists, markets, governance structures, and social norms. These decisions, and the institutional practices they support, compose an important subject for moral evaluation and can be assessed from the standpoint of distributive justice.

A key element of liberal political theory, as articulated by Rawls, is the distinction between the "basic structure" of society, which is subject to principles of distributive justice, and other domains of life that are not directly subject to these principles. The basic structure encompasses

> the way in which the major social institutions fit together into one system, and how they assign fundamental rights and duties and shape the division of advantages that arise through social cooperation. Thus the political constitution, the legally recognized forms of property, and the organization of the economy, and the nature of the family, all belong to the basic structure.[6]

These practices need to be structured in accordance with a common set of rules. Outside of these contexts, people are left relatively free to pursue their personal objectives, something that is important for a pluralistic society in which people have divergent goals and aspirations.

Against this backdrop, I wish to advance two claims. The first is that the basic structure of society is best understood as a composite of sociotechnical systems: that is, systems that are constituted through the interaction of human and technological elements. The claim here is not only that the basic structure contains social and technical elements, but also that these elements interact dynamically to constitute new forms of stable institutional practice and behavior.[7] The second is that AI increasingly shapes elements of the basic structure in relevant ways, and hence that its design, development, and deployment all potentially interface with principles of justice in this context.

The growing role played by AI in the operation of key institutions and practices is well illustrated by the criminal justice system, in which risk-assessment algorithms increasingly determine a person's eligibility for bail or parole, facial recognition technology has been used to augment police capabilities, and AI systems direct the allocation of policing resources using predictive analytics. In the context of economic mobility and access to key public services such as welfare provision, the use of algorithmic tools is similarly influential, determining who is eligible for welfare support, who has access to public housing, and which families are engaged by child services.[8] Meanwhile, in the economic sphere, financial institutions use these models to determine who has access to loans, mortgages, and insurance. Finally, these tools have a wider impact on the economic prospects of citizens via their integration into job recommendation search engines – helping to determine who is shown what opportunities – and via the tools used by educational institutions to allocate students or advertise opportunities for higher education.[9]

In each case, AI is not simply an additional ingredient that supervenes onto a stable practice leaving the fundamental elements of that practice untouched. Rather, AI interacts with the behavior of human decision-makers to shape the

character of these practices, including how they distribute benefits and burdens across the population. In the context of criminal justice, for example, there is significant concern that parole recommendation algorithms compound historical injustice by recreating and extending racial bias found in the training data.[10] In the context of government services, AI has changed the nature of welfare provision, including who can access it and on what terms, with political scientist Virginia Eubanks documenting the emergence of a "feedback loop of injustice" whereby "marginalized groups face higher levels of data collection when they access public benefits . . . [which] acts to reinforce their marginality when it is used to target them for suspicion and extra scrutiny."[11] Meanwhile, in the domain of credit scoring and access to financial services, legal scholar Frank Pasquale has raised concerns about the increasingly significant role played by a person's algorithmically determined "digital reputation" as a major determinant of their life chances.[12] Speaking to the dynamic interaction between these systems and the social environment in which they are deployed, Pasquale notes that "unlike the engineer, whose studies do nothing to the bridges she examines, a credit scoring system increases the chance of a consumer defaulting once it labels him a risk and prices a loan accordingly."[13] Given the potential serious knock-on effects these practices have for equality at the societal level, they have driven concerns about "digital redlining" – with entire groups of people encountering new barriers to opportunity – and the emergence of what, with respect to race, sociologist Ruha Benjamin terms "the New Jim Code."[14]

To be clear, the concerns that arise in these contexts are not only concerns about distributive justice, they also involve racial justice, criminal justice, historic injustice, and the disciplinary power of institutions.[15] However, principles of distributive justice that spell out how major institutions ought to allocate opportunities and resources are also relevant here. Moreover, they can help explain what is morally problematic about these practices and show how these harms can be addressed.

According to the Rawlsian framework, there are two key grounds that make a practice subject to regulation by principles of distributive justice, both of which are now met by the aforementioned AI systems. First, these principles apply to institutions that are necessary in order to maintain "background justice" over time.[16] According to this view, a social practice should be regulated by principles of distributive justice when, without this intervention, the compound effect of individual choices would lead to forms of inequality that threaten the equal standing and autonomy of citizens.[17] For example, the uninterrupted interplay of market forces would likely leave some people so badly off that they could no longer give meaningful consent to the institutional practices that structure their lives, and would instead have to accept whatever arrangement was

offered to them by the rich and powerful. To avoid this outcome, the practices that make up the basic structure need to be regulated in ways that support background justice, counteracting the tendency of multiple individual transactions to distort the distribution of income and wealth over time.

What is important for our purpose is that in modern societies, background justice is increasingly mediated algorithmically. Across various contexts, including social service provision, credit allocation, and insurance eligibility decisions, AI systems have now taken on this critical function. By making assessments or predictions based upon an individual's past choices, and by providing decisions or recommendations that then shape that person's opportunity set, these systems exert a strong influence on the unfolding relationship between individual choices and collective outcomes. Moreover, unless their operation is aligned with principles of distributive justice, these systems could compound inequality in ways that a just society aims to forestall.

Second, principles of distributive justice apply to certain practices because they exercise a "profound and pervasive impact" upon a person's life chances.[18] In particular, they shape the terms on which people can access the benefits of social cooperation, the development of their personal goals and aspirations, and the occasions on which they encounter the coercive power of the state. For many AI systems, this threshold is now also being met. In the words of legal scholar Rashida Richardson, AI systems are now being used to determine

> who will have their food subsidies terminated, how much healthcare benefits a person is entitled to, and who is likely to be a victim of crime .... [They] have concrete consequences for individuals and communities, such as increased law enforcement harassment, deportation, denial of housing or employment opportunities, and death.[19]

The stakes are therefore sufficiently high for principles of justice to be invoked.

If the preceding argument is correct, then it has a number of implications for the character of AI systems that are deployed in these spaces. These include:

Publicity. The theory of justice developed by Rawls aims to identify principles for the governance of major institutions that can be justified to people despite variation in their beliefs about what a good or perfect society would look like. Situated in the "original position," people are asked to choose principles of justice for society from behind a "veil of ignorance," which prevents them from knowing the position in society they will occupy. Given that people are not able to tailor principles in a way that is prejudicial to their own interests, the principles selected are held to be fair and thus ones that people can willingly endorse. Moreover, given that society at times relies upon coercive sanctions to enforce norms via legal instruments, Rawls holds that "the grounds of its institutions should stand up to public scrutiny."[20] This "publicity condition" ensures that "citizens are in a po-

sition to know and to accept the pervasive influences of the basic structure that shape their conception of themselves, their character and their ends."[21]

The publicity condition has important ramifications for the uses of AI that we have discussed. In particular, the requirement appears to sit in tension with elements of what Pasquale terms the "black box society," including the use of opaque hiring and credit allocation algorithms that shape citizen's life prospects.[22] Conversely, it helps to explain why calls for certain kinds of explanation are justified in the context of these AI systems: they are part of a more general entitlement citizens hold in relation to the institutions that shape their lives.[23] Moreover, as we have seen, mere knowledge of the principles that govern the behavior of public institutions is not sufficient to render them legitimate. People must also be in a position to accept the principles despite variation in personal moral beliefs. In the context of AI, this means that the integration and deployment of the technology must be justifiable in terms of an ideal of public reason.[24] It should, in the words of philosopher Jonathan Quong, be something that is acceptable "to each of us by reference to some common point of view, despite our deep differences and disagreements."[25]

One major consequence of this requirement is that an appeal to purely private goals, whether those of an individual or organization, will not be sufficient to justify the adoption or deployment of AI systems in certain public contexts. Instead, a public rationale must be provided. Second, the publicity condition points toward the existence of a derivative duty on the part of those who develop and deploy AI systems – to test them prior to deployment and to offer nontechnical explanations of their performance – so that the models are amenable to this kind of informed public debate, discussion, and evaluation.

Basic liberties. The first principle of justice endorsed by Rawls requires that "each person has the same indefeasible claim to a fully adequate system of basic liberties, which scheme is compatible with the same scheme of liberties for all."[26] These basic liberties work to "protect fundamental interests that have special significance" and include, at a minimum, "freedom of thought and liberty of conscience; the political liberties and freedom of association, as well as the freedoms specified by the liberty and integrity of the person; and finally, the rights and liberties covered by the rule of law."[27] The basic liberties are relevant for the design and deployment of AI systems in at least two respects.

The first concerns the protection they accord citizens. A major aim of this principle is to ground "a secure common status of equal citizenship" for society's members.[28] This aspiration dovetails effectively with the notion that institutions must be "effectively and impartially administered," given that deviation from this ideal contravenes the rights and liberties covered by the rule of law.[29] Understood in this way, the enjoyment of equal basic liberties stands in opposition to certain forms of algorithmic discrimination. As philosopher Tommie Shelby notes, the principle prohibits cases in which the rules of a public institution are applied un-

evenly, including situations "where the administration or enforcement of its rules and procedures is frequently distorted by the racial prejudice and bias of its officials."[30] While the primary concern at the time of writing was with the bias of human officials, there is no reason to think that bias is less problematic when it is inherited by automated decision systems that perform a similar function. Indeed, given the potential for these systems to perform better than human decision-makers, one might think that the errors they make are more egregious.

Second, the list of basic liberties provided by Rawls is dynamic and varies according to the sociotechnical character of the society to which they apply. The initial list is based upon conditions that are held to be necessary for the development of moral autonomy and personhood irrespective of time or place (such as freedom of conscience). However, Rawls also notes that it is wise to take a "historical approach," which involves identifying additional rights that have demonstrable practical value for different societies at a specific moment in time. As a consequence, Rawls writes that "it is perhaps impossible to give a complete specification of these liberties independent from the particular circumstances – social, economic and *technological* – of a given society."[31] On each occasion, the key question is: what liberties are necessary to protect individuals in the development and pursuit of the conception of the good life, given the specific sociotechnical character of the society in which they live?

The potential for intrusion created by modern AI systems, both in terms of the data they are trained on and their ability to influence or foreshadow subsequent behavior, has given range to a host of new concerns.[32] To guard against these risks, it is quite possible that a right to privacy should now be added to the list of basic liberties. Although the grounds of a potential right to privacy are philosophically contested, legal scholar Andrei Marmor argues that they are closely connected to our well-being and are "violated when somebody manipulates, without adequate justification, the relevant environment in ways that significantly diminish your ability to control what aspects of yourself you reveal to others."[33] Given Rawls's concern with the ability of citizens to pursue a conception of the good life that is free from unwarranted interference, the basic liberties may now include protection against invasive forms of surveillance or behavioral manipulation.

Fair equality of opportunity. Rawls's second principle of justice holds that:

> Social and economic inequalities are to satisfy two conditions: first they are to be attached to offices and positions open to all under conditions of fair equality of opportunity; and second, they are to be to the greatest benefit of the least-advantaged member of society.

This principle also has far-reaching implications for AI. Starting with the first condition, it holds that fair equality of opportunity – not just formal equality of opportunity – must be achieved when determining how opportunities are allocat-

ed between citizens. Thus, the requirements of justice are not met simply via the adoption of processes that do not discriminate against people on the basis of certain protected characteristics at the point at which a decision is made. Instead, a just society will aim to eliminate the impact of a wide range of unchosen features on their life prospects. The most natural reading of this requirement includes features such as a person's race, sex, class, and other contingencies of birth. Once the relevant adjustments have been made, we should arrive at a situation in which people of similar ability have roughly equal prospects of success.

In the context of debates around AI fairness, the implications of this principle are potentially significant. They mean moving away from a purely formal conception of fairness as equal treatment or "de-biasing" and thinking about how these tools can actively mitigate the effect of bias that exists at a societal level through various corrective measures.[34] As information scientists Solon Barocas and Andrew Selbst have noted, this debate mirrors a long-running discussion in jurisprudence about the appropriate goal of antidiscrimination legislation.[35] Whereas the anticlassification approach is concerned with equal treatment in a formal sense that involves eliminating unfairness that "individuals in certain protected classes experience due to decision makers' choices," antisubordination reaches beyond that and is more closely aligned with Rawls's fair equality of opportunity principle.[36] It holds that the goal of antidiscrimination law is "to eliminate status-based inequality due to membership in those classes, not as a matter of procedure, but of substance."[37] If this is the appropriate normative standard for AI systems performing key social functions, then we need further research and public discussion about what substantively fair outcomes look like in practice, and about how AI systems can support this societal objective.

The difference principle. The second condition, commonly known as the *difference principle*, also has implications for the design and deployment of AI. This principle holds that for institutional practices to be just, all inequalities in the distribution of "social primary goods" (which include income, wealth, and the "social bases of self-respect") must work to the greatest advantage of the least advantaged member of society. It follows that when AI is integrated into a key social practice, in a way that affects the overall distribution of benefits and burdens, it is pertinent to ask whether it does the most it possibly can do to improve the position of the least advantaged member of that system. This is a challenging question, and one that points toward a potentially exacting standard for AI deployment. Cumulatively, it redraws the scope of current debates about how to evaluate the impact of AI systems, making the impact of these systems on the distribution of wealth, resources, and social standing an important desideratum, while also proposing a standard for evaluation that is strongly egalitarian.

In terms of the practical implications of the difference principle, it seems clear that any technology that worsens the position of the most disadvantaged mem-

ber of society in absolute terms, once it has been incorporated into relevant social practice, will fail to meet a key requirement of justice irrespective of other benefits it may bring (such as scalability or efficiency). Yet fully realized, the difference principle proposes a higher standard than simply improving the status quo: it suggests that the AI systems must make the worst-off as well-off as they can, relative to alternative system designs, or otherwise risk being part of a practice that is not fully legitimate. This standard is most clearly applicable to AI systems that have been integrated into core economic functions. However, it potentially has much wider applicability, extending to the full range of sociotechnical systems that shape a person's access to resources or impact upon their social standing and sense of self-worth.

Moreover, this demand is not met simply by the present combination of private innovation in the space of AI and *post hoc* economic redistribution. For while the redistribution of wealth is an important component of justice on any account, we also need to consider how sociotechnical systems influence the production of inequality *ex ante*. This is because there are likely to be opportunities to intervene at this point that do not arise later on. Indeed, given the emphasis Rawls places on self-esteem, in particular, there are opportunities to create fairer AI systems (that minimize inequalities in the first place) that cannot be addressed simply by making those who are badly off the *post facto* recipient of wealth transfers. Ultimately, these opportunities are what is missed when technology is bracketed out from liberal political theory: we may fail to consider an important site of distributive justice and hence mistakenly believe that society is substantively just when this is not the case and when impermissible forms of technologically induced inequality are hiding in plain sight.

These arguments are presented in the spirit of constructive co-investigation. My main purpose has been to illustrate the kind of rich moral insight that results from extending the domain of distributive justice to include AI systems. Clearly, more work needs to be done to substantiate these claims and translate them into guidelines for technologists and public officials. Indeed, as this preliminary account makes clear, it is possible that tensions will emerge, for example, between the notion derived from the liberty principle that individuals must be treated in a consistent manner and the notion, anchored in the fair equality of opportunity principle, that groups must experience similar outcomes.[38] Nonetheless, core elements of this approach seem destined to remain in place. If AI is, as I have argued, now part of the major sociotechnical practices that make up the basic structure of society, then its design and deployment should feed into practices that are amenable to public justification, support citizen's rights, and embody substantive properties connected with an egalitarian conception of justice. In these contexts, the appropriate goal of AI alignment is not an open ques-

tion. Rather, the development and deployment of AI systems represent a new site for the operation of principles of distributive justice.[39]

I have argued that when AI is integrated into the functioning of major institutions and social practices, the norms that apply to the basic structure of society also apply to these systems. To ground this claim, I pointed to the role that AI now plays in augmenting or undermining background justice, and to a range of profound effects that AI has on the lives of citizens, particularly in the context of our major political and economic institutions. However, the preceding argument leaves open the question of alignment for AI systems deployed outside of key socioeconomic practices. In these environments, is it perhaps the prerogative of engineers or organizations to align AI systems with their own preferred values?

To answer this question, we need to understand how the two grounding conditions map onto other kinds of AI systems. Taking the profound effects condition first, it seems likely that many AI deployments meet this threshold. For example, AI-powered search and curation systems are deeply integrated into prevailing social epistemological practices, functioning as custodians for the legibility of the world around us, and influencing what we take to be true on an individual and collective level. Moreover, recommendation systems have the potential to influence the development of our moral character in certain ways, shaping self-perception, preferences, and desires, even as they learn to "give us what we want." Yet when it comes to background justice the case for an expansive reading is less clear. As we have seen, background justice is concerned with society's ability to reproduce itself over time in such a way that the conditions for meaningful consent are preserved. From this vantage point, certain forms of interpersonal exploitation and domination are clearly objectionable. The salient question for AI systems is whether there are other roles they play, beyond those considered, that also mandate corrective measures of this kind.

Given uncertainty on this point, efforts to extend principles of distributive justice to a wider set of AI systems are somewhat inconclusive. Yet even on a restrictive reading of the scope of these principles, two further points remain to be made. First, principles of distributive justice have implications for AI systems that are not part of the basic structure. On this point, Rawls notes that we should not regard the "political and the nonpolitical domains as two separate, disconnected spaces . . . each governed solely by its own distinct principles."[40] Instead, principles of justice place "essential restrictions" on all other activities. By way of illustration, Rawls does not consider the media to be part of the basic structure of society. However, requirements of justice nonetheless entail that this sphere of activity must be structured in a way that ensures the fair value of the political liberties. In the context of AI, it means that, at a minimum, public deployments of this technology must be compatible with principles of justice. Moreover, on an individual

level, liberal political theory holds that we are all under a "duty of justice" to support the operation of institutions that enable cooperation on terms that are fair. When applied to groups concerned with the creation of new technologies, duties of justice plausibly become "duties of deployment" to support, and not subvert, the functioning of just institutions.

Second, the demand for public justification in the context of AI deployment may well extend beyond the basic structure. As social scientist Langdon Winner argues, when the impact of a technology is sufficiently great, this fact is, by itself, sufficient to generate a free-standing requirement that citizens be consulted and given an opportunity to influence decisions.[41] Absent such a right, citizens would cede too much control over the future to private actors, something that sits in tension with the idea that they are free and equal. Against this claim, it might be objected that it extends the domain of political justification too far, in a way that risks crowding out room for private experimentation, exploration, and the development of projects by citizens and organizations. However, the objection rests upon the mistaken view that autonomy is promoted by restricting the scope of justificatory practices to as narrow a subject matter as possible. In reality, this is not the case: what matters for individual liberty is that practices that have the potential to interfere with this freedom are appropriately regulated so that infractions do not come about. Understood in this way, the demand for public justification stands in opposition not to personal freedom but to forms of unjust technological imposition.[42]

The demand for justice in the context of AI is well-founded. Considered through the lens of distributive justice, key principles that govern the fair organization of our social, political, and economic institutions also apply to AI systems that are embedded in these practices. One major consequence of this is that liberal and egalitarian norms of justice apply to AI tools and services across a range of contexts. When they are integrated into society's basic structure, these technologies should, I have argued, support citizens' basic liberties, promote fair equality of opportunity, and provide the greatest benefit to those who are worst-off. Moreover, deployments of AI outside of the basic structure must still be compatible with the institutions and values that justice requires. There will always be valid reasons, therefore, to consider the relationship of technology to justice when it comes to the deployment of AI systems.

AUTHOR'S NOTE

I would like to thank Laura Weidinger, William Isaac, Julia Haas, Conor Griffin, Sean Legassick, Christopher Summerfield, Allan Dafoe, Shakir Mohamed, Brittany Smith, Courtney Biles, Aliya Ahmad, Geoff Keeling, Thomas K Gilbert, Abeba Birhane, Jeff Howard, Juri Viehoff, Johannes Himmelreich, James Manyika, and the editorial team at *Dædalus* for their support with this work.

ABOUT THE AUTHOR

Iason Gabriel is a Staff Research Scientist at DeepMind. He has published in such journals as *Minds and Machines*, *The Philosophical Quarterly*, and *The Journal of Applied Philosophy*.

ENDNOTES

[1] Matthew Le Bui and Safiya Umoja Noble, "We're Missing a Moral Framework of Justice in Artificial Intelligence," in *The Oxford Handbook of Ethics of AI*, ed. Markus D. Dubber, Frank Pasquale, and Sunit Das (Oxford: Oxford University Press, 2020), 163.

[2] John Rawls, *A Theory of Justice*, rev. ed. (Cambridge, Mass.: Harvard University Press, 1999), 130.

[3] Peter Dietsch, "G. A. Cohen, Karl Marx's Theory of History: A Defence," in *The Oxford Handbook of Classics in Contemporary Political Theory*, ed. Jacob T. Levy (Oxford: Clarendon Press, 2000).

[4] Key elements of this approach have been defended by Lewis Mumfold, Langdon Winner, Andrew Feenberg, Ursula Franklin, Ruha Benjamin, and Jeroen Van Der Hoven, among others.

[5] Ruha Benjamin, *Race After Technology: Abolitionist Tools for the New Jim Code* (Hoboken, N.J.: John Wiley & Sons, 2019), 2.

[6] John Rawls, *Political Liberalism* (New York: Columbia University Press, 2005), 258.

[7] Wiebe E. Bijker, *Of Bicycles, Bakelites, and Bulbs: Toward a Theory of Sociotechnical Change* (Cambridge, Mass.: MIT Press, 1997), 273–274.

[8] Virginia Eubanks, *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor* (London: Picador, 2018).

[9] D. J. Pangburn, "Schools Are Using Software to Help Pick Who Gets In. What Could Go Wrong?" *Fast Company*, May 17, 2019; and Manish Raghavan and Solon Barocas, "Challenges for Mitigating Bias in Algorithmic Hiring," Brookings Institution, December 6, 2019.

[10] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner, "Machine Bias," ProPublica, May 23, 2016, 139–159; and Kristian Lum and William Isaac, "To Predict and Serve?" *Significance* 13 (5) (2016): 14–19.

[11] Eubanks, *Automating Inequality*, 6–7.

[12] Frank Pasquale, *The Black Box Society* (Cambridge, Mass.: Harvard University Press, 2016), 14.

[13] Ibid., 41.

[14] Safiya Umoja Noble, *Algorithms of Oppression* (New York: New York University Press, 2018); and Benjamin, *Race After Technology*.

[15] Indeed, given the breadth of these concerns, rejection of the argument contained in this essay does not entail that concerns about justice in the context of AI are not valid. They could still be grounded in other ways. See, for example, Sasha Costanza-Chock, *Design Justice: Community-Led Practices to Build the Worlds We Need* (Cambridge, Mass.: MIT Press, 2020).

[16] John Rawls, *Justice as Fairness: A Restatement* (Cambridge, Mass.: Harvard University Press, 2001), 10.

[17] Miriam Ronzoni, "The Global Order: A Case of Background Injustice? A Practice-Dependent Account," *Philosophy & Public Affairs* 37 (3) (2009): 229–256.

[18] Rawls, *A Theory of Justice*, 82.

[19] Rashida Richardson, "Defining and Demystifying Automated Decision Systems," *Maryland Law Review* (forthcoming): 1.

[20] Rawls, *Political Liberalism*, 68.

[21] Ibid.

[22] Pasquale, *The Black Box Society*.

[23] Kate Vredenburgh, "The Right to Explanation," *Journal of Political Philosophy* (2021).

[24] See Reuben Binns, "Algorithmic Accountability and Public Reason," *Philosophy & Technology* 31 (4) (2018): 543–556; and Thomas Krendl Gilbert, "Mapping the Political Economy of Reinforcement Learning Systems: The Case of Autonomous Vehicles," *Simons Institute Newsletter*, January 31, 2021.

[25] Jonathan Quong, "Public Reason," *The Stanford Encyclopedia of Philosophy*, May 20, 2013, updated October 24, 2017, https://plato.stanford.edu/archives/spr2018/entries/public-reason/.

[26] Rawls, *Justice as Fairness*, 42.

[27] Rawls, *Political Liberalism*, 291.

[28] Rawls, *A Theory of Justice*, 199.

[29] Ibid., 48.

[30] Tommie Shelby, "Race and Social Justice: Rawlsian Considerations," *Fordham Law Review* 72 (5) (2004): 1706.

[31] Rawls, *A Theory of Justice*, 54 (italics mine).

[32] Carissa Véliz, *Privacy Is Power: Why and How You Should Take Back Control of Your Data* (New York: Random House, 2020).

[33] Andrei Marmor, "What Is the Right to Privacy?" *Philosophy & Public Affairs* 43 (1) (2015): 3–26.

[34] Ben Green, "Escaping the 'Impossibility of Fairness': From Formal to Substantive Algorithmic Fairness," arXiv (2021), https://arxiv.org/abs/2107.04642.

[35] Solon Barocas and Andrew D. Selbst, "Big Data's Disparate Impact," *California Law Review* 104 (2016): 671.

[36] Ibid.

[37] Ibid., 723.

[38] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan, "Inherent Trade-Offs in the Fair Determination of Risk Scores," arXiv (2016), https://arxiv.org/abs/1609.05807.

[39] For important complementary analysis of the way in which Rawlsian considerations have influenced discussion of information ethics, see Anna Lauren Hoffman, "Beyond Distributions and Primary Goods: Assessing Applications of Rawls in Information Science and Technology Literature Since 1990," *Journal of the Association for Information Science and Technology* 68 (7) (2017): 1601–1618.

[40] Rawls, *Justice as Fairness*, 166.

[41] Langdon Winner, *The Whale and the Reactor: A Search for Limits in an Age of High Technology* (Chicago: University of Chicago Press, 2020), 9.

[42] Iason Gabriel, "Artificial Intelligence, Values, and Alignment," *Minds and Machines* 30 (3) (2020): 411–437.

Artificial Intelligence, Humanistic Ethics

John Tasioulas

Ethics is concerned with what it is to live a flourishing life and what it is we morally owe to others. The optimizing mindset prevalent among computer scientists and economists, among other powerful actors, has led to an approach focused on maximizing the fulfilment of human preferences, an approach that has acquired considerable influence in the ethics of AI. But this preference-based utilitarianism is open to serious objections. This essay sketches an alternative, "humanistic" ethics for AI that is sensitive to aspects of human engagement with the ethical often missed by the dominant approach. Three elements of this humanistic approach are outlined: its commitment to a plurality of values, its stress on the importance of the procedures we adopt, not just the outcomes they yield, and the centrality it accords to individual and collective participation in our understanding of human well-being and morality. The essay concludes with thoughts on how the prospect of artificial general intelligence bears on this humanistic outlook.

Ethics is, first and foremost, a domain of ordinary human thought, not a specialist academic discipline. It presupposes the existence of human choices that can be appraised by reference to a distinctive range of values. The delimitation of this range, among other values such as aesthetic or religious values, is philosophically controversial. But on a fairly standard reading, two very general, interlocking questions lie at the heart of ethics: What is it to live a good or flourishing life? And what is it that we owe to others, notably fellow human beings, but also nonhuman animals or even inanimate nature? The first question brings us into the territory of individual well-being; the second into that of morality, especially the obligations we owe to others and the rights they hold against us. Philosophers expound theories of well-being and morality and their interrelations, but all of us, in living our lives, constantly make choices that reflect answers to these questions, however inchoate or unconscious they may be.

Engagement with ethics is inescapable in decision-making about artificial intelligence.[1] The choices we make regarding the development and deployment of AI-based technologies are ultimately intelligible only in terms of the fallible pursuit of ethical values such as the acquisition of knowledge and control or the promotion of health, justice, and security. Moreover, all forms of "regulation" that might be proposed for AI, whether voluntary self-regulation in deciding whether

© 2022 by John Tasioulas
Published under a Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) license
https://doi.org/10.1162/DAED_a_01912

to use a social robot as a caregiver, or the social and legal norms that should govern the manufacturing and use of such robots, ultimately implicate choices that reflect judgments about ethical values and their prioritization.

A clear-eyed appreciation of the pervasive significance of ethics for AI is sometimes obscured by an odd contraction that the idea of ethics is liable to undergo in this domain. So, for example, Kate Crawford, author and founder of the AI Now Institute, urges us to "focus less on ethics and more on power" because "AI is invariably designed to amplify and reproduce the forms of power it has been deployed to optimize."[2] But what would the recommended focus on power entail? For Crawford, it means interrogating the institutional power structures in which AI is embedded by reference to ideas of equality, justice, and democracy. But the irony is that these three ideas are either themselves core ethical values or, in the case of democracy, need to be explicated and defended in terms of such values.

Nonetheless, Crawford's injunction usefully prompts reflection on the various ways the idea of ethics has been unduly diminished in recent discussions about AI, no doubt partly a result of the prominent role of big tech players in shaping the field of "AI ethics" to limit the threat it poses to their commercial ambitions. Consider three ways the diminishment of ethics is typically effected.

Content. The content of ethical standards is often interpreted as exclusively a matter of fairness, which is primarily taken to be a relational concern with how some people are treated compared with others. Illustrations of AI-based technology that raise fairness concerns include facial recognition technology that systematically disadvantages darker-skinned people or automated resume screening tools that are biased against women because the respective algorithms were trained on data sets that are demographically unrepresentative or that reflect historically sexist hiring practices. "Algorithmic unfairness" is a vitally important matter, especially when it exacerbates the condition of members of already unjustly disadvantaged groups. But this should not obscure the fact that ethics also encompasses nonrelational concerns such as whether, for example, facial recognition technology should be deployed at all in light of privacy rights or whether it is disrespectful to job applicants in general to rank their resumes by means of an automated process.[3]

Scope of application. Ethics is sometimes construed as narrowly individualistic in focus: that is, as being concerned with guiding individuals' personal conduct, rather than also bearing on the larger institutional and social settings in which their decisions are made and enacted.[4] In reality, however, almost all key ethical values, such as justice and charity, have profound implications for institutions and patterns of social organization. Plato's *Republic*, after all, sought to understand justice in the individual soul by considering it "writ large" in the polity. Admittedly, some philosophers treat political justice as radically discontinuous from justice in the soul. The most influential proponent of the discontinuity thesis in

recent decades is John Rawls, who contends that pervasive reasonable disagreement on ethical truth disqualifies beliefs about such truths from figuring as premises in political justification.[5] This is a sophisticated controversy, which cannot be addressed here, save to note that this kind of move will always face the response that the phenomenon of reasonable disagreement, and the need for respect that it highlights, is itself yet a further topic for ethical appraisal, and hence cannot displace the need to take a stand on ethical truth.[6]

Means of enforcement. There is a widespread assumption that ethics relates to norms that are not properly enforceable – for example, through legal mechanisms – but instead are backed up primarily by the sanction of individual conscience and informal public opinion. But the general restriction of ethics to "soft" forms of regulation in this way is arbitrary. The very question whether to enact a law or other regulatory norm and, if so, how best to implement and enforce it, is one on which ethical values such as justice and personal autonomy have a significant bearing. Indeed, there is a long-standing tradition, cutting across ideological boundaries, that identifies justice precisely with those moral rights that should in principle receive social and legal enforcement.

In short, we should reclaim a broad and foundational understanding of ethics in the AI domain, one that potentially encompasses deliberation about any form of regulation, from personal self-regulation to legal regulation, and which potentially has radical implications for the reordering of social power.

Given its inescapability, ethical thought is hardly absent from current discussions around AI. However, these discussions often suffer from a tendency either to leave inexplicit their operative ethical assumptions or else to rely upon them uncritically even when they are made explicit. We can go even further and identify a dominant, or at least a prominent, approach to ethics that is widely congenial to powerful scientific, economic, and governmental actors in the AI field.

Like anyone else, AI scientists are prone to the illusion that the intellectual methods at their disposal have a far greater problem-solving purchase than is warranted. This is a phenomenon that Plato diagnosed in relation to the technical experts of his day, artisans such as cobblers and shipbuilders. The mindset of scientists working in AI tends to be data-driven, it places great emphasis on optimization as the core operation of rationality, and it prioritizes formal and quantitative techniques. Given this intellectual orientation, it is little wonder that an eminent AI scientist, like Stuart Russell, in his recent book *Human Compatible: AI and the Problem of Control,* is drawn to preference-based utilitarianism as his overarching ethical standpoint.[7]

Russell's book takes the familiar worry that AI – in the form of an artificial general intelligence (AGI) that surpasses human intellectual capabilities – will even-

tually spiral out of control, unconstrained by human morality, with disastrous consequences. But what is human morality? Russell appears to take it as axiomatic that the morally right thing to do is whatever will maximize the fulfilment of human preferences.[8] In terms of our two core concerns of ethics, the fulfilment of human preferences is taken to encompass well-being, and the fundamental moral injunction is to maximize overall well-being thus conceived. So ethics is reduced to an exercise in prediction and optimization: which act or policy is likely to lead to the optimal fulfilment of human preferences?

But this view of ethics is notoriously open to multiple serious – I believe, fatal – objections. Its concern with aggregating preferences threatens to override important rights that erect strong barriers to what can be done to individuals. Why not feed a few Christians to the lions if their preferences to stay alive are outweighed by the preferences of a sufficiently large number of blood-thirsty Roman spectators? And that is even before we observe that many preferences are infected with racism, sexism, or other prejudices; that they may reflect false or incomplete information; or that they may be psychological adaptations to oppressive circumstances. Ethics operates in the crucial space of reflection on what our preferences should be, a vital consideration that makes a belated appearance in the last few pages of Russell's book.[9] It cannot take those preferences as ultimate determinants of value.

There are moral philosophers who defend versions of preference utilitarianism that are patched-up to address these difficulties. But the idea that preference utilitarianism is a highly contestable moral theory does not really register in Russell's book, which conforms with my suspicion that it approximates to a default position among leading actors in the AI field.

The same broad approach is heavily influential among leading economic and governmental actors. This is perhaps less obvious, since the doctrine is standardly modified by positing wealth-maximization as the more readily measurable proxy for preference satisfaction. Hence the tendency of GDP to hijack governmental decision-making around economically consequential technologies, with the resultant sidelining of values that are not readily catered to by the market, such as public goods like access to justice and health care or the preservation of a sustainable environment. Hence, also, the legitimation of profit maximization by corporations as the most effective institutional means to societal wealth maximization. Of course, many who adopt such an approach have never heard of utilitarianism or, if they have, may explicitly reject it. But one revealing indication of the dominance of an ideology is the way that people who disavow it can nonetheless remain in its intellectual grip.

A key priority for those working in the field of AI ethics is to elaborate an ethical approach that transcends the limitations and distortions of this dominant ethical paradigm. In my view, such a humanistic ethics – one

that encompasses aspects of human engagement with the ethical that are not adequately captured by the methods of natural science and mainstream economics, but that are the traditional concern of the arts and humanities – would possess at least the following three, interrelated features (the three Ps).

Pluralism. The approach would emphasize the plurality of values, both in terms of the elements of human well-being (such as achievement, understanding, friendship, and play) and the core components of morality (such as justice, fairness, charity, and the common good). This pluralism of values abandons the comforting notion that the key to the ethics of AI will be found in a single master concept, such as trustworthiness or human rights. How could human rights be the comprehensive ethical framework for AI when, for example, AI has a serious environmental impact beyond its bearing on anthropocentric concerns? And what of those important values to which we do not have a right, such as mercy or solidarity? Nor can trustworthiness be the master value. Being parasitic on compliance with more basic values, trustworthiness cannot itself displace those values.

Beyond the pluralism of values is their incommensurability. We are often confronted with practical problems that implicate an array of values that pull in different directions. In such cases, although some decisions will be superior to others, there may be no single decision that is optimal: in choosing an occupation, teaching may be a better field for me than surgery, but we cannot assume there is a single profession that is, all things considered, best, rather than a limited array of eligible alternatives that are no worse than the others. This incommensurability calls into question the availability of some optimizing function that determines the single option that is, all things considered, most beneficial or morally right, the quest for which has animated a lot of utilitarian thinking in ethics.

It is worth observing that confidence about the deployment of AI to minimize "noise" in human judgment – the unwanted variability, for example, in hiring decisions by employers or sentencing by judges – displayed in the important new work of Daniel Kahneman, Olivier Sibony, and Cass Sunstein, sometimes involves an implicit reductionism about the values at stake that downplays the scope for incommensurability.[10] For example, the authors treat bail decisions fundamentally as predictions of the likelihood that the accused will abscond or reoffend, sidelining considerations such as the gravity of the offense with which they have been charged or the impact of detention on the accused's dependents.[11] But such decisions typically address multivalue problems, and there is no guarantee that there is a single best way of reconciling the competing values in each case. This means not only that algorithms will need to be more sophisticated to balance multiple salient values in reaching a correct decision, but that much of what looks like noise may be acceptable variability of judgments *within* the range of rationally eligible alternatives.

Procedures, not only outcomes. Of course, we want AI to achieve valuable social goals, such as improving access to education, justice, and health care in an effective and efficient way. The COVID-19 pandemic has cast into sharp relief the question of what outcomes AI is being used to pursue: for example, is it enabling physicians to diagnose and triage patients faster and more effectively, or is it primarily engaged in profit-making activities, like vacuuming up people's attention online, that have little or no redeeming social value?[12] The second feature of a humanistic approach to ethics emphasizes that what we rightly care about is not just the value of the outcomes that AI applications can be used to deliver, but the procedures through which it does so.

If, for example, important practical decisions exhibit the phenomenon of incommensurability, then we may have good reason to ensure that they are assigned to humans, rather than to automated processes, to preserve a valuable form of autonomy for humans as they express and develop their tastes and characters in choosing from divergent, but rationally eligible, pathways in life. Of course, there is the further question of how to balance such autonomy against demands for consistency (or "noiselessness"), especially in public decision-making. Should we tolerate significant divergence in sentencing across judges, or should the demands for "horizontal equity" prevail, ensuring that like cases are treated alike? Proponents of the latter view often recommend the use of algorithms to guide or replace human decision-making. This itself is a difficult question of striking a balance between competing considerations in our legal culture, with no *ex ante* guarantee that one solution will emerge as superior overall.

But the case for according ultimate decision-making authority to humans can also be made even if we suppose that a single correct answer is always available. Take, for example, the use of AI in cancer diagnosis and its use in the sentencing of criminals. Intuitively, the two cases seem to exhibit a difference in the comparative valuing of the soundness of the eventual decision or diagnosis and the process through which it is reached. When it comes to cancer, generating the most accurate diagnosis may be all-important, it being largely a matter of indifference whether this is generated by an AI diagnostic tool or the exercise of human judgment. In criminal sentencing, however, being sentenced by a robot judge – even if the sentence is likely to be less biased or less "noisy" than one rendered by a human counterpart – appears to sacrifice important values, such as the ideal of reciprocity among fellow citizens that is central to the rule of law.[13]

This last point is familiar, of course, in relation to such process values as transparency, procedural fairness, and explainability. Even if the procedure followed by the judicial algorithm can be made transparent, there is a serious question – given, for example, the vast discrepancy between machine learning and ordinary human reasoning processes – whether it affords an explanation of the right kind, an explanation that a criminal defendant can grasp as offering intelligible reasons for

the decision to imprison him. But the point goes beyond the important issue of explainability. How does it feel to contemplate the prospect of a world in which judgments that bear on our deepest interests and moral standing have, at least as their proximate decision-makers, autonomous machines that do not have a share in human solidarity and cannot be held accountable for their decisions in the way that a human judge can?

Participation. The third feature relates to the importance of participation in the process of decision-making with respect to AI, whether as an individual or as part of a group of self-governing democratic citizens. At the level of individual well-being, this takes the focus away from theories that equate human well-being with an end state such as pleasure or preference-satisfaction. These end states could in principle be brought about through a process in which the person who enjoys them is passive: for example, by the government putting a happiness drug into the water supply. Contrary to this passive view, it would stress that successful engagement with valuable pursuits is at the core of human well-being.[14]

If the conception of human well-being that emerges is deeply participatory, then this bears heavily on the delegation of decision-making power to AI applications. One of the most important sites of participation in constructing a good life, in modern societies, is the workplace.[15] According to a McKinsey study, around 30 percent of all work activities in 60 percent of occupations could one day be automated.[16] Can we accept the idea that the large-scale elimination of job opportunities can be compensated for by the benefits that automation makes available? The answer partly depends on whether the participatory self-fulfilment of work can, any time soon and for the vast majority of those rendered jobless, be feasibly replaced by other activities, such as art, friendship, play, or religion. If it cannot, addressing the problem with a mechanism like a universal basic income, which involves the passive receipt of a benefit, will hardly suffice. Instead, much greater attention will need to be paid to how AI can be integrated into productive practices in ways that do not so much replace human work as enhance its quality, making it more productive, fulfilling, and challenging, while also less dangerous, repetitive, and lacking in meaning.[17]

Similarly, we value citizen participation as part of collective democratic self-government. And we do so not just because of the instrumental benefits of democratic decision-making in generating superior decisions by harnessing cognitive diversity, but also because of the way in which participatory decision-making processes affirm the status of citizens as free and equal members of the community.[18] This is an essential plank in the defense against the tendency of AI technology to be co-opted by technocratic modes of decision-making that erode democratic values by seeking to convert matters of political judgment into questions of technical expertise.[19]

At present, much of the culture in which AI is embedded is distinctly technocratic, with decisions about the "values" encoded in AI applications being taken by corporate, bureaucratic, or political elites, often largely insulated from meaningful democratic control. Indeed, a small group of tech giants accounts for the lion's share of investment in AI research, dictating its overall direction and setting the prevalent moral tone. Meanwhile, AI-enabled social media risks eroding the quality of public deliberation that a genuine democracy needs, such as by promoting the spread of disinformation, aggravating political polarization, or using bots in astroturfing campaigns. Similarly, the use of AI as part of corporate and governmental efforts to monitor and manipulate individuals undermines privacy and threatens the exercise of basic liberties, effectively discouraging citizen participation in democratic politics.[20]

As with workplace participation, we need to reflect seriously on how AI and digital technology more generally can enable, rather than hinder and distort, democratic participation.[21] This is especially urgent given the declining faith in democracy across the globe in recent years, including in long-established democracies such as the United Kingdom and the United States. Indeed, the disillusionment is such that, in a recent poll, 51 percent of Europeans favored replacing at least some of their parliamentarians with AI.[22] There is still time to salvage the democratic ideal that an essential part of civic dignity is participation in self-government.

An additional complexity here concerns how these two modes of participation – in the workplace and in politics – are connected. It is obvious that active participation in the two domains is mutually reinforcing in important ways. Thus, powers of reason and sociability that are developed in a participatory workplace, and that foster a sense of equal civic dignity, can be brought to bear in democratic deliberation about political questions, just as democratic control over the impact of new technologies on the workplace can help preserve and enhance its vital role as a site of genuine human fulfilment.[23]

I have mainly focused on narrow AI, conceived as AI-powered technology that can perform limited tasks (such as facial recognition or medical diagnosis) that typically require intelligence when performed by humans. This is partly because serious doubt surrounds the likelihood of artificial general intelligence emerging within any realistically foreseeable time frame, partly because the operative notion of "intelligence" in discussions of AGI is problematic,[24] and partly because a focus on AGI often distracts us from the more immediate questions of narrow AI.[25]

With these caveats in place, however, one can admit that thought experiments about AGI can help bring into focus two questions fundamental to any humanistic ethic: What is the ultimate source of human dignity, understood as the inherent value attaching to each and every human being? And how can we relate hu-

man dignity to the value inhering in nonhuman beings? Toward the end of Kazuo Ishiguro's novel *Klara and the Sun*, the eponymous narrator, an "Artificial Friend," speculates that human dignity – the "human heart" that "makes each of us special and individual" – has its source not in something within us, but in the love of others for us.[26] But a threat of circularity looms for this boot-strapping humanism, for how can the love of others endow us with value unless those others already have value? Moreover, if the source of human dignity is contingent on the varying attitudes of others, how can it apply equally to every human being? Are the unloved bereft of the "human heart"?

Questions like these explain the tendency among some to interpret the inherent value of each individual human being as arising from the special love that a supremely good transcendent being – God, represented by the sun, in Ishiguro's novel, which the solar-powered Klara treats as a kind of life-sustaining divinity – has for each human being in equal measure.[27] But invoking a divine being to underwrite human dignity leads us into obvious metaphysical and ethical quagmires, which in turn raise the difficult question of whether the inherent worth of human beings can be explicated within a broadly naturalistic framework.[28] Supposing that it can be, this is compatible with a distinct kind of dignity also inhering in other beings, such as nonhuman animals.

We are still struggling to integrate the value of nonhuman animals within our ethical thought. Doing so requires overcoming the baleful influence of longstanding practices in which animals are treated either as possessing merely instrumental value in relation to human ends, or at best intrinsic value that is conditional on their role in human life. The dream of AGI, should it ever become a reality, will generate an even more acute version of this problem, given the prominent role that our rational capacities play in elevating human dignity above the dignity of other beings known to us.[29] For the foreseeable future, however, our focus must be on properly integrating AI technology into a culture that respects and advances the dignity and well-being of humans, and the nonhuman animals with whom we share the world, rather than on the highly speculative endeavor of integrating the dignity of intelligent machines into our existing ethical framework.

AUTHOR'S NOTE

This essay began life as a blog post for the Ada Lovelace Institute, "The Role of the Arts and Humanities in Thinking about Artificial Intelligence (AI)." I am grateful to the Institute for permitting me to reuse some material here. I have benefited from comments on previous drafts from Dominic Burbidge, Hélène Landemore, Seth Lazar, Ted Lechterman, James Manyika, Adrian Vermuele, Carina Prunkl, Divya

Siddarth, Carissa Veliz, Glen Weyl, Mike Woolridge, and John Zerilli. I regret that I have not been able to pursue many of their very stimulating comments within the confines of this short essay.

ABOUT THE AUTHOR

John Tasioulas is Professor of Ethics and Legal Philosophy at the Faculty of Philosophy and Director of the Institute for Ethics in AI at the University of Oxford. He is the editor of *The Cambridge Companion to the Philosophy of Law* (2020) and *The Philosophy of International Law* (with Samantha Besson, 2010).

ENDNOTES

[1] I shall assume a very broad understanding of AI as essentially the use of machines to perform tasks that characteristically require intelligence when performed by humans. My focus will primarily be on "narrow" AI applications, such as facial recognition, surveillance, and risk-assessment, rather than artificial general intelligence, though I say something about the latter toward the very end.

[2] Kate Crawford, *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence* (New Haven and London: Yale University Press, 2021), 224.

[3] Some of these issues are compellingly developed by Joshua Cohen in "Don't Shoot the Algorithm" (unpublished manuscript).

[4] The "effective altruism" movement, which has significant allegiance among tech elites, is arguably one expression of this depoliticized and, in its effect, ultimately conservative view of ethics. See Amia Srinivasan, "Stop the Robot Apocalypse," *London Review of Books* 37 (18) (2015).

[5] John Rawls, *Political Liberalism* (New York: Columbia University Press, 1993). For an attempt to pursue the radical discontinuity thesis in relation to AI, see Iason Gabriel, "Artificial Intelligence, Values, and Alignment," *Minds and Machines* 30 (3) (2020): 411.

[6] See John Tasioulas, "The Liberalism of Love," in *Political Emotions: Towards a Decent Public Sphere*, ed. Thom Brooks (London: Palgrave Macmillan, forthcoming 2022).

[7] I am here identifying an influential mode of thought that Russell's book epitomizes. It should be emphasized, however, that there have always been scientists in this domain who have urged the importance of a multidisciplinary approach with an important humanistic dimension, such as in Joseph Weizenbaum, *Computer Power and Human Reason: From Judgment to Calculation* (San Francisco: Freeman & Co, 1976); and, more recently, in Nigel Shadbolt and Roger Hampson, *The Digital Ape: How to Live (in Peace) with Smart Machines* (London: Scribe, 2018).

[8] Stuart Russell, *Human Compatible: AI and the Problem of Control* (London: Allen Lane, 2019), 178.

[9] Ibid., 255.

[10] Daniel Kahneman, Olivier Sibony, and Cass R. Sunstein, *Noise: A Flaw in Human Judgment* (London: William Collins, 2021).

[11] Ibid., chap. 10.

[12] For a discussion of studies showing that AI predictive tools made no real difference in diagnosing and triaging COVID-19 patients, and in some cases, may have been harmful, see William Douglas Heaven, "Hundreds of AI Tools Have Been Built to Catch Covid, None of Them Helped," *MIT Technology Review*, July 30, 2021, https://www.technologyreview.com/2021/07/30/1030329/machine-learning-ai-failed-covid-hospital-diagnosis-pandemic/.

[13] John Tasioulas, "The Rule of Law," in *The Cambridge Companion to the Philosophy of Law*, ed. John Tasioulas (Cambridge: Cambridge University Press, 2020), 131–133.

[14] Joseph Raz, *The Morality of Freedom* (Oxford: Oxford University Press, 1986), chap. 12.

[15] Anca Gheaus and Lisa Herzog, "The Goods of Work (Other Than Money!)," *Journal of Social Philosophy* 47 (1) (2016): 70–89.

[16] James Manyika and Kevin Sneader, "AI, Automation, and the Future of Work: Ten Things to Solve For," McKinsey Global Institute Executive Briefing, June 1, 2018, https://www.mckinsey.com/featured-insights/future-of-work/ai-automation-and-the-future-of-work-ten-things-to-solve-for.

[17] For an exploration of this theme, see Frank Pasquale, *New Laws of Robotics: Defending Human Expertise in the Age of AI* (Cambridge, Mass.: Harvard University Press, 2020).

[18] For a powerful recent defense of democracy along these lines, see Josiah Ober, *Demopolis: Democracy Before Liberalism in Theory and Practice* (Cambridge: Cambridge University Press, 2017).

[19] For a candid statement, by a Silicon Valley billionaire, of the need to harness the libertarian promise of technology to an antidemocratic ethos, see Peter Thiel, "The Education of a Libertarian," *Cato Unbound: A Journal of Debate*, April 13, 2009, https://www.cato-unbound.org/2009/04/13/peter-thiel/education-libertarian.

[20] For a helpfully wide-ranging discussion of some of these issues, see Joshua Cohen and Archon Fung, "Democracy and the Digital Public Sphere," in *Digital Technology and Democratic Theory*, ed. Lucy Bernholz, Hélène Landemore, and Rob Reich (Chicago: University of Chicago Press, 2021).

[21] For some positive thinking along these lines, see Hélène Landemore, "Open Democracy and Digital Technologies," in *Digital Technology and Democratic Theory*. For useful discussions of digitally enhanced democracy in pioneering countries such as Estonia and Taiwan, see Hans Kundani, *The Future of Democracy in Europe: Technology and the Evolution of Representation* (London: Chatham House, 2020), https://www.chathamhouse.org/sites/default/files/CHHJ7131-Democracy-Technology-RP-INTS-200228.pdf; and Divya Siddarth, *Taiwan: Grassroots Digital Democracy That Works* (New York: Radical Exchange, 2021), https://www.radicalxchange.org/media/papers/Taiwan_Grassroots_Digital_Democracy_That_Works_V1_DIGITAL_.pdf.

[22] "More Than Half of Europeans Want to Replace Lawmakers with AI, Study Finds," CNBC, May 27, 2021, https://www.cnbc.com/2021/05/27/europeans-want-to-replace-lawmakers-with-ai.html. For an interesting discussion of "algocracy" ("rule by algorithm"), see Ted Lechterman, "Will AI Make Democracy Obsolete?" *Public Ethics*, August 4, 2021, https://www.publicethics.org/post/will-ai-make-democracy-obsolete. It is worth noting that proposals for algocracy often assume that the point of politics is

to aggregate human preferences, in line with the preference-based utilitarianism discussed above.

[23] For a perceptive discussion of the way AI threatens to disrupt the economic underpinnings of democracy, see Daron Acemoglu, *Redesigning AI: Work, Democracy, and Justice in the Age of Automation* (Boston: Boston Review Forum, 2021).

[24] Like many others, Stuart Russell adopts an impoverished conception of intelligence as competence in means-ends reasoning according to which the choice of ends is made extraneous to the operations of intelligence. On this view, a machine that annihilated humanity in order to maximize the number of paper clips in existence can qualify as superintelligent. Ibid., 167. For a wide-ranging and perceptive discussion of problems with the idea of "intelligence" invoked in discussions of AGI, see Divya Siddarth, Daron Acemoglu, Danielle Allen, et al., "How AI Fails Us" (Cambridge, Mass.: Edmond J. Safra Center for Ethics, 2021).

[25] Some of these concerns are discussed in John Tasioulas, "First Steps Towards an Ethics of Robots and Artificial Intelligence," *Journal of Practical Ethics* 7 (1) (2019): 61–95, http://www.jpe.ox.ac.uk/papers/first-steps-towards-an-ethics-of-robots-and-artificial-intelligence/.

[26] Kazuo Ishiguro, *Klara and the Sun* (London: Faber, 2021), 218, 306.

[27] Nicholas Wolterstorff, *Justice: Rights and Wrongs* (Princeton, N.J.: Princeton University Press, 2008), 352–361.

[28] David Wiggins, *Solidarity and the Root of the Ethical* (Lawrence: The Lindley Lecture, University of Kansas, 2008).

[29] See, for example, John Tasioulas, "Human Dignity and the Foundations of Human Rights" in *Understanding Human Dignity*, ed. Christopher McCrudden (Oxford: Oxford University Press, 2013), 293–314; and Jeremy Waldron, *One Another's Equals: The Basis of Human Equality* (Cambridge, Mass.: Harvard University Press, 2017).

Automation, Augmentation, Value Creation & the Distribution of Income & Wealth

Michael Spence

Digital technologies are transforming the economy and society. The dimensionality and scope of the impacts are bewildering and too numerous to cover in a single essay. But of all the concerns around digital technology (and there are many), perhaps none has attracted more attention, and generated deeper anxiety, than the impact of various types of automation on work and on the structure of the economy. I focus on the ways in which the digitization of virtually all data, information, and content is transforming economies. And more specifically, I look at the impacts of automation, augmentation, AI, machine learning, and advanced robotics on economic transformations, on work, and on the distribution of income and wealth.

Digital technology can be thought of as digital machines (computers, servers, and various other portable devices), software, and networks (with standardized protocols) creating, storing, operating on, and transmitting information in digital form. "Use cases" refer to applications of digital technology such as mobile payments, social media, online commerce, and location-specific services like maps.

Automation and digital machine augmentation are a class of use cases. *Automation* involves replacing people with machines in the performance of certain tasks that machines can carry out without human intervention or guidance. *Augmentation* connotes adding machines to a work environment, enabling people to be more productive.[1] The two As are flip sides of the same coin.

Augmentation of human productivity using machines is hardly new. If you think of tools as simple machines, augmentation of human capabilities has characterized most of human history on the planet. In the first industrial revolution, when energy and power were added to the mix via steam engines, and later electricity and fossil fuels, machine augmentation produced (with a lag) a huge, sustained acceleration in productivity. We called it mechanization. And it changed work, raising concerns that there would not be enough jobs to go around, or that a subset of people would not be able to learn how to do the new jobs that required working with machines.

© 2022 by Michael Spence
Published under a Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) license
https://doi.org/10.1162/DAED_a_01913

This kind of machine augmentation can produce transitory unemployment for a few reasons. First, work in the sectors experiencing rapid machine augmentation requires new sets of skills, and these do not adjust immediately in the workforce. Second, utilizing the new technology often requires the installation of new production systems and business models, which again is far from instantaneous. And third, increased productivity will give rise to rising incomes, but that may not result in higher demand right away. For instance, at present, labor incomes in many countries are diverging from productivity for much of the population; income is going to capital and toward the wealthy end of the income spectrum, where savings are higher.

That said, longer term, two parallel processes have typically prevented permanent unemployment from becoming a reality. Incomes rise, and with that increase comes elevated demand for goods and services. Second, with higher incomes, labor markets adapt to different work-leisure trade-offs; hours worked steadily decline over long periods of time. Data on hours worked support this proposition. Across countries, hours worked per year declines with per capita income, and across time among Organisation of Economic Co-operation and Development countries, hours worked declines over time.[2] (See Figure 1.) In the pre-industrial era, many people worked long hours just to provide for basic needs.

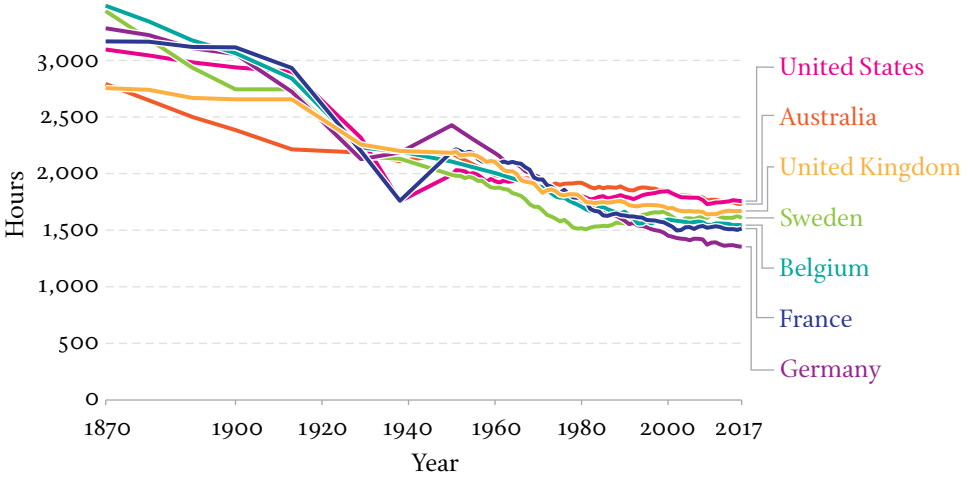
Nevertheless, concerns about employment in the aggregate – in effect, whether there are enough jobs to go around – have been common in periods of rapid technological advancement.

In the predigital era, the issue was not about what might be called *full automation*, because machines did not perform tasks all by themselves. Mechanization was the focus of attention. In modern parlance, one can think of it as machine augmentation, but not digital machines. People used machines to do tasks much more quickly, often with higher-quality outputs, and even to produce things that were impossible in the premachine age.

Admittedly, the historical line between mechanization and automation is sometimes a little blurry. The history of weaving machines or looms contains fascinating examples of predigital automation.[3] But for the most part, in the predigital era, machines did not carry out long and relatively complex sequences of actions without human intervention. In general, machines augmented and replaced humans in the physical performance of tasks, but the information and control layer, governing the sequencing and timing of the activities of machines, remained firmly in human hands or minds. Machines in the industrial age were powerful but they did not function autonomously. In the digital era, this is no longer true.

This somewhat simple fact helps explain why we find ourselves in uncharted territory. In any economy there is a very large collection of activities that involve gathering, recording, analyzing, and transferring information, transactions, coordination of activity, pattern recognition, and decisions. There is an enormous

Figure 1
Annual Average Working Hours per Worker



Note: Before 1950, the data correspond only to full-time production workers (non-agricultural activities). Starting in 1950, estimates cover total hours worked in the economy as measured from primarily National Accounts data. Our World in Data plotted the data from Huberman and Minns and extended coverage using an updated vintage of the Penn World Table, which uses the same underlying source. Comparisons between countries are limited due to differences in measurement. Source: Michael Huberman and Chris Minns, "The Times They Are Not Changin': Days and Hours of Work in Old and New Worlds, 1870–2000," *Explorations in Economic History* 44 (4) (2007); and University of Groningen, Penn World Table 9.1 database (2019).

amount of embedded knowledge in an economy, and sophisticated transmission mechanisms via institutions and informal networks that support information flows and decisions (large and small) that in the aggregate determine the performance of the economy.

Which brings us to automation, digital machines, and software. Early computers were programmed manually to perform a specific function, with wires, somewhat like an old telephone switchboard.[4] That cumbersome and limiting technology was quickly replaced by programmable computers, which load a set of digital coded instructions and then execute them autonomously. The code contains the logical sequence of steps that would be carried out by humans were they performing the task. Up until recent advances in artificial intelligence (over the past decade), this has been the basic model of digital automation.

Before the programmable computer and the digitization of most information in the form of data, automation and augmentation in the information, control, coordination, decision, and transactions layers of the economy were negligible. Now, as of about the past sixty years, for the first time we have powerful machines in the information, coordination, and decision (ICD) layer. That is in part why digital economists Erik Brynjolfsson and Andrew McAfee referred to this as the *second machine age* in their influential book of the same name.[5]

For those who might have suspected that this ICD layer was perhaps an important but relatively minor (in terms of value-added) part of the economy, the effect of the COVID-19 pandemic on the economy was revealing. Substantially disconnecting the functioning of the ICD layer from the need for physical proximity, which is what in part modern network-based digital technology has done, enabled economies to keep operating with substantially reduced mobility and physical contact. Of course, there are important sectors in which proximity is unavoidable that either contracted or shut down during the pandemic or stayed open but at considerable health risk to their workers.

In short, the ICD layer, which is effectively the governing and control mechanism in the economy and the market system, now has powerful machines that automate, replace, and sometimes outperform people in some tasks, while simultaneously augmenting people in the performance of other tasks.

It is tempting to assume that the second machine age will follow a pattern like industrialization, and indeed it may. But as we are in somewhat uncharted territory, we should not rush to this conclusion. While the impact of digital technology has already been substantial, the technology seems clearly set to continue to advance.

The full economic impact of AI and machine learning, for example, is largely in the future. The pandemic economy has accelerated digital adoption across a range of sectors, including many that had been lagging in this respect.[6] Powerful machine learning tools are now widely available for a rapid digital transformation.

The first round of digital automation involved codifying tasks. Codifiable tasks have two properties: 1) people can do them – meaning, carry out the steps and 2) we can figure out and precisely describe the logical steps we use in performing them, steps and logic that are then embodied in code. It is the second part that enables digital machines to automate significant parts of the information-processing and control layer of the economy. It is also the part that blocked progress in artificial intelligence until the machine learning revolution, because a variety of "tasks" – such as image recognition, understanding natural language, translation, and in fact a number of pattern-recognition and prediction capabilities that humans have – defied codification. This impasse in AI put quite severe limits on the scope for automation.

AI based on machine learning broke this impasse essentially with an end run. Take image recognition, an area that has experienced dramatic advances in the

past ten years. Instead of trying to find a set of "rules" for classifying images, deep learning algorithms analyze millions of digital images to detect patterns. Over time, the predictions about images become increasingly accurate. Similar advances are occurring in speech, language translation, a host of medical applications, and many more fields.

For the economy, the significance of the breakthroughs in AI via machine learning lies in dramatically increasing the scope for automation and, critically, augmentation across the entire economy. Advanced robotics, autonomous vehicles, radiology, analysis of DNA sequences in relation to diseases, reading vast amounts of literature for doctors and other professionals, expanding access to credit and other financial services via algorithms that close informational gaps are all enabled by machine learning. And these are just a few examples.

The first round of automation, now in its middle age, was not AI-driven but required codification of tasks. We have data on its impacts. This second round of automation and augmentation enabled by machine learning is in its early stages. Its full impact is not yet apparent in economic data.

The impact of the codification version of automation has already been substantial. Routine white- and blue-collar jobs, or more precisely jobs in which routine tasks are a large component, began declining with a noticeable acceleration around the year 2000. Here, routine refers to jobs that substantially consist of tasks that are codifiable and hence subject to partial or complete automation. Codifiable, however, does not mean simple. The sequence of steps including conditional branching, true/false determinations, and classification can be long and complex.

Since many of these routine jobs were associated with mid-level incomes, the first order effect in economic terms has been to reduce employment in routine, middle-income jobs. A distinctive characteristic of digital automation is that it includes both blue- and white-collar work, the latter involving processing, storing, recording, and retrieving information.

A resulting pattern of job and income polarization has been documented for most of the developed economies.[7] The immediate effect on the income distribution is to flatten it, making the tails larger and the middle smaller. Of course, if the displacement of middle-income jobs were extreme, we could start to see bimodal distributions, but that has not yet happened.

How does the economy respond when a certain class of jobs is automated? Initially, there may be transitory unemployment, and adjustment typically takes place dynamically as follows. At first, people will look for jobs that match their current specific skill set. But that does not work for everyone because those jobs are declining in the economy. They then set employment as a priority and move to jobs in which the skill requirements are either fewer, or are more easily acquired

quickly, as we have seen with part-time jobs in the Internet-enabled gig economy. Often that means lower incomes. What one sees from a macroeconomic perspective is not necessarily large-scale increases in unemployment, but rather a deterioration in the income distribution. One can think of this as a shifting equilibrium, assuming the skills on the supply side remain fixed.[8]

However, that is not the end of the story, because the skills and human capital side of the job market are not fixed, just slower moving. People look for and start investing in skills and human capital that are in demand in nonroutine job categories with higher incomes. The pace of this more time-consuming process crucially depends on the presence or absence of supporting institutions, including employers. That is why, for example, there are numerous partnership initiatives in the United States, involving government at all levels, businesses, and educational institutions to accelerate the skills-transition process. And in many other countries, the existing and well-developed educational and skills-training infrastructure is being adapted to digital transformations to continue to play a key role in these supply-side skills transitions.

A second key gatekeeping factor with respect to the skills transition is the distribution of income and wealth. Investing in one's own human capital takes time and financial resources regardless of the quality of the institutional support mechanisms. If income and wealth inequality is extreme, then the lower part of the income/wealth distribution will struggle to make the investments in their own human capital, the more so in the absence of high-quality publicly funded and low-cost key public services.

Notice the circularity here. Automation of middle-income jobs has contributed to suppressing the middle of the income distribution. For those who are pushed toward lower deciles, the challenge of investing their way out via new skills acquisition becomes more difficult. So the income distribution is both an outcome of and an input to the digital transitions in work. The appropriate conclusion seems to be that policies that directly address high income inequality will turn out to contribute to successful work transitions, even if that is not the primary purpose of the policies.[9]

The skills-adjustment process can have a beneficial effect on the income distribution. Essentially, it partially undoes the adverse initial distributional effect of automation by increasing the supply of people with skills that are in high demand and have higher incomes. In fact, other things equal, it may lower the incomes in these higher-skill/income segments, or the so-called skills premium. But it also reduces the supply of people in the lower-skill/income part of the spectrum, and hence puts upward pressure on the wages there.

A natural and important question is, "Will the skills-adjustment process largely eliminate the adverse initial distributional effects of automation?" An honest answer, I think, is that we do not know because we have no way of determining

with any precision, ultimately, how fungible the human capital stock really is in the population. The existing evidence does not support the view that adverse distributional effects are transitory. To date, skill-biased technical change (substantially in the digital area) looks like it has shifted income upward in the distribution, even after the skills-adjustment process on the supply side is well underway.[10]

The main takeaways here are: 1) we are grappling with complex structural changes and transitions in work, skill requirements, and human capital, not equilibria; 2) the purpose and end point of this transition is to turn automation into digital augmentation; 3) technology is not stationary with the result that the target keeps moving, especially with the application of machine learning across most sectors of the economy; and 4) extreme income inequality combined with institutional and policy shortfalls risk turning a complex transition into a trap for the lower-income part of the population.

Two other dimensions of the challenge of AI and automation deserve attention. One is that technology can and does adapt in ways that reduce the magnitude of the skills-acquisition problem. We see this all around us. Digital equipment and systems are designed to be easier to use, and markets reward that kind of innovation. Perhaps the best historical example is the graphical user interface (now found on virtually every consumer digital device). It is so pervasive that we all take it for granted and for younger people, it is simply the normal way to interact with digital machines. But there was a time when interacting with "computers" was a lot less intuitive, and largely confined to those with the requisite training.

Let me turn now to AI and machine learning technologies. As noted earlier, machine learning implies a vast expansion in the tasks or subtasks that can be automated: for example, advanced robotics, autonomous vehicles, and reading and editing technical literature for, say, medical professionals. In addition, machine learning–based pattern-recognition applications go well beyond human capabilities in some areas, as applied to genetics and biomedical science, for example, taking it well beyond automation and firmly into the realm of augmentation.[11]

Since these advances in AI for work entail a significant expansion of tasks or subtasks that can be automated and performed by machines, one can ask where on the income spectrum these work-related disruptions will land. Again, an honest answer is that it is too soon to know with any confidence. But a reasonable guess is across the board, and not mainly in the middle-income range of white- and blue-collar work, as in the first round of code-based digital automation. At one end, low-income labor-intensive manufacturing in developing countries is set to be disrupted in a way that was not possible before the recent advances in AI and robotics, and that presents challenges to the development model of low-income countries. On the other end, scientific research and technology development and

high-end professional services look to be within the target range as well, not in the sense of full automation, but rather in substantial digital machine augmentation, via the automation of key tasks.

Thus far, we have focused mainly on various aspects of the digital transformation in developed economies. But their impact does not stop there. There are at least two important classes of developing economies in which the impact of automation and augmentation has large and diverse current and future impacts: middle-income countries (often called emerging economies) and the lower-income countries in which the growth and development process is in the early stages.

Emerging economies are developing countries that have reached middle-income levels. Some are growing quickly while others struggle with growth. But for the most part, they have resources and, in general, reasonably well-developed digital infrastructure. But they still have poor segments of the population in need of informational and related services. Many live outside major urban areas with limited access to traditional offline services. In these economies, probably more than either developed or lower-income economies, the digital transformations are a large net positive in multiple dimensions. Generally, these economies have left behind the labor-intensive manufacturing and assembly growth and employment engines that are threatened by modern AI and advanced robotics and that represent a significant obstacle for lower-income countries seeking to replicate the high-growth patterns achieved by their predecessors.

In the emerging economies, e-commerce, mobile payments, and fintech – all now powered by AI – are closing the service availability gaps associated with remoteness. The same is true in education and health care. In addition, the lower-income parts of the population have traditionally had difficulty accessing modern services because of the absence of documentable identities and financial track records. I call this the anonymity problem. Digital data and machine learning are proving to be powerful tools for bridging these gaps and overcoming the obstacles.

And on the employment front, the middle-income economies are in complex transitions in which the service sectors are expanding as a share of the economy and employment. Generally, the digital transformations are accelerating these changes and creating considerably more jobs than are being eliminated. This is not to say the previous discussion of digital transformations and work are entirely different here. There are skills transitions to be navigated, too. But the balance is different. It is more about training than retraining. Job losses associated with automation are small in relation to the job losses in the declining labor-intensive sectors. Indeed, automation is seen as a way to keep manufacturing sectors that are transitioning from labor-intensive to digitally capital-intensive. These are sectors that, in the absence of the productivity increases that go with automation, would migrate to lower-income parts of the global economy. In addition, legacy systems

that tend to hold back the pace of change in developed economies are less developed in emerging economies. So the new digitally based economy, as it emerges from the advances in technology and use cases, is in many ways embraced faster. Technologically, they are leapfrogging intermediate steps.

There is also a very interesting trend in global entrepreneurial activity that is directly relevant to employment and structural change in emerging economies: a rapid increase in new company formation across a range of emerging economies and continents. It is linked in large part to the rapid spread of the digital economy. The financial and other parts of innovative ecosystems that support this have also become global in coverage. The result is high-growth companies and unicorns (privately held startup companies with a value of over $1 billion) proliferating across the globe. (See Table 1.)

As this trend gains momentum, it provides powerful new employment engines and entrepreneurial opportunities, especially for younger parts of the population. The opportunity is created in part by automation in the development and improved performance of new markets and complementary systems. These systems are increasingly architected, modified, and improved by AI. This dimension of automation is much less discussed than automation in the context of work, but it is an important element of AI-powered automation and augmentation. At least in the digital area, the entry barriers are low and the initial capital requirements are also low, making them accessible, ideal environments for fostering innovation, entrepreneurship, and new company formation.

For the lower-income countries, the picture is similar in some respects but different in others, presenting a challenge and an opportunity. The mobile Internet has substantially closed the gap in terms of basic digital infrastructure, though there is more investment needed to bring coverage, network capacity, speed, and reliability up to middle-income-country levels. For these economies, a growth model is needed that leverages global economy demand and technology. Traditionally, for non-resource-rich countries, the core of the growth model has been exports of labor-intensive goods and process-oriented manufacturing and assembly. It is a model based on low labor costs and has historically had the virtue of being a powerful employment engine.

The problem is that AI and advanced robotics are making inroads into the growth and employment model that depends on low-cost labor, in which the comparative advantage lay in labor-intensive manufacturing (think textiles and apparel). This aspect of automation is making labor costs much less influential in determining where to locate production. It is leading to what economist Dani Rodrik has termed premature deindustrialization. The problem is that it is thus far not clear whether there is an equally capable alternative engine to power growth and development.

Table 1
Entrepreneurship Is Now a Global Phenomenon

Global Unicorns: Top Ten Countries by Number of Unicorns					
Country	**Number of Unicorns**	**Percent of Total**	**Total Valuation ($Billions)**	**Percent of Total**	**Top Three Unicorns by Value**
United States	337	52%	$1,093	51%	Stripe, SpaceX, Instacart
China	138	21%	$547	26%	Bytedance, DiDi Chuxing, Yuanfudao
India	31	5%	$106	5%	One97 Communications, BYJU's, OYO Rooms
United Kingdom	29	4%	$96	5%	Checkout.com, Global Switch, Hopin
Germany	16	2%	$29	1%	Otto Bock Healthcare, N26, Celonis
Israel	13	2%	$17	1%	Earnix, Monday.com, Wiz
Brazil	12	2%	$42	2%	Nuback, Wildlife Studios, Loft
South Korea	10	2%	$22	1%	Krafton, Yello Mobile, Toss
France	10	2%	$13	1%	BlaBlaCar, Alan, Mirakl
Canada	5	1%	$11	1%	PointClickCare, Dapper Labs, Clearco
Top Ten Countries	601	92%	$1,976	92%	
Total Unicorns	654		$2,147		

Note: As of April 2021, there are now more than 600 unicorns around the world, predominantly led by the United States (337 unicorns) and China (138 unicorns). China, India, and Brazil stand out in this picture. There is good reason to believe that countries like Indonesia and many others will make an appearance soon. Source: CB Insights, "Entrepreneurship Is Now a Global Phenomenon," April 2021.

That said, the benefits in terms of digitally enabled inclusive growth patterns, described above for emerging economies, including especially expanded opportunity for entrepreneurial activity, also apply to the lower-income countries, provided the digital infrastructure is in place. The importance of this should not be understated for fostering inclusive growth patterns.

How you think about these issues depends on the unit of analysis. If the unit is a working person, then automation can eliminate the job. More commonly, though, the result is augmentation: it replaces part of the job, changing the nature of the work. On the other hand, if you start with the unit being some subsystem of the economy, say a manufacturing facility, then machines are essentially augmentation, just as they were in the industrial revolution, and automation, which is new and digitally enabled, is an important and powerful tool in making us more productive.

ABOUT THE AUTHOR

Michael Spence, a Fellow of the American Academy since 1983, is Senior Fellow at the Hoover Institution, and Dean Emeritus and the Philip H. Knight Professor Emeritus of Management in the Graduate School of Business at Stanford University. He was awarded the Nobel Prize in Economics in 2001. He is the author of *The Next Convergence: The Future of Economic Growth in a Multispeed World* (2012) and *Market Signaling: Informational Transfer in Hiring and Related Screening Processes* (1974).

ENDNOTES

[1] Productivity in this context is broader than the standard economic concept. To be sure, workers will become more productive with digital machines, but so too will biomedical scientists, armed with powerful AI tools, or software engineers working with computers that can code.

[2] "List of Countries by Average Annual Labor Hours," Wikipedia, last updated November 23, 2021, https://en.wikipedia.org/wiki/List_of_countries_by_average_annual_labor_hours.

[3] Melinda Watt, "Weaving Machinery," Love to Know, https://fashion-history.lovetoknow.com/fashion-clothing-industry/weaving-machinery. The Jacquard loom, which automated the production of patterned fabrics for the first time in 1904, used punch cards to set the pattern for raising the warp fibers, before the weft spindle crossed over.

[4] An early computer could be wired to do arithmetic calculations and it worked quite well. But if you wanted to change the calculation, you had to physically rewire the device, much like an early manual telephone switching device.

[5] Erik Brynjolfsson and Andrew McAfee, *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies* (New York: W. W. Norton & Co., 2014).

[6] Studies by the McKinsey Global Institute of digital adoption and digital footprints across sectors of the economy show both advanced sectors and many lagging sectors. One of the anticipated effects of the pandemic is accelerated adoption, especially in the lagging sectors. See, for example, Susan Lund, Anu Madgavkar, James Manyika, et al., *The Future of Work after COVID-19* (New York: McKinsey Global Institute, 2021), https://www.mckinsey.com/featured-insights/future-of-work/the-future-of-work-after-covid-19.

[7] David Autor, "Why Are There Still So Many Jobs? The History and Future of Workplace Automation," *Journal of Economic Perspectives* 29 (3) (2015).

[8] It is possible for labor-saving and skill-biased technological shifts to leave labor overall worse off. See Anton Korinek and Joseph E. Stiglitz, "Artificial Intelligence, Globalization, and Strategies for Economic Development," Institute for New Economic Thinking Working Paper Series No. 146 (New York: Institute for New Economic Thinking, 2021), https://doi.org/10.36687/inetwp146.

[9] We have seen this kind of pattern before. Another major force operating on the structure of developed economies has been globalization and international specialization. For an analysis of how that impacted the economy and jobs across the economy, see Michael Spence and Sandile Hlatshwayo, "The Evolving Structure of the American Economy and the Employment Challenge," *Comparative Economic Studies* 54 (4) (2012): 703–738.

[10] There is nothing in economic theory that assures that the distribution of income that results from the operation of market and technological forces is socially acceptable. If one adds to the mix declining bargaining power for lower-income parts of the labor force, and the influence of wealth on policies that affect the distribution of income and access to key services in education and health, that reinforces the conclusion. So the skills-adjustment process is important and deserves the attention it is getting, but it is not the whole story with respect to restoring broadly inclusive growth patterns.

[11] OpenAI has created AI systems that write computer code, and that can translate code from one computer language to another. DeepMind in London, a subsidiary of Alphabet, has algorithms that can predict with reasonable accuracy the 3D structures of proteins, a crucial step in drug and vaccine development and other aspects of biomedical research.

Automation, AI & Work

Laura D. Tyson & John Zysman

We characterize artificial intelligence as "routine-biased technological change on steroids," adding intelligence to automation tools that substitute for humans in **physical** *tasks and substituting for humans in routine and increasingly nonroutine* **cognitive** *tasks. We predict how AI will displace humans from existing tasks while increasing demand for humans in new tasks in both manufacturing and services. We also examine the effects of AI-enabled digital platforms on labor. Our conjecture is that AI will continue, even intensify, automation's adverse effects on labor, including the polarization of employment, stagnant wage growth for middle- and low-skill workers, growing inequality, and a lack of good jobs. Though there likely will be enough jobs to keep pace with the slow growth of the labor supply in the advanced economies, we are skeptical that AI and ongoing automation will support the creation of* **enough good jobs***. We doubt that the anticipated productivity and growth benefits of AI will be widely shared, predicting instead that they will fuel more inequality. Yet we are optimistic that interventions can mitigate or offset AI's adverse effects on labor. Ultimately, how the benefits of intelligent automation tools are realized and shared depends not simply on their technological design but on the design of intelligent policies.*

Amazing new automation and digital technologies are transforming work and the economy.[1] Artificial intelligence is the latest tool in a toolkit of "automation" technologies that perform tasks previously performed by humans, often more cheaply, faster, and better.

Most humans depend on income from work for their livelihoods, and we focus on how AI, like other forms of automation, affects work. A key question is how AI and AI-enabled intelligent tools will impact the supply of and access to good jobs that provide middle-class earnings, safe working conditions, legal protections, social insurance and benefits, and career-building opportunities. In the advanced market economies and democracies that are the focus of this essay, political and social stability depends on the availability and accessibility of good jobs.

With supportive fiscal and monetary policies adequate to maintain high levels of employment, it is likely that there will be *enough jobs* to keep pace with the slow growth of the labor supply in these economies. But we are skeptical that AI along with ongoing automation will support the creation of *enough good jobs*. And we are doubtful that the anticipated productivity and growth benefits of AI will be

© 2022 by Laura D. Tyson & John Zysman
Published under a Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) license
https://doi.org/10.1162/DAED_a_01914

widely shared, predicting instead that they will fuel income and wealth inequality. Yet we are optimistic that wise interventions can change the trajectory of AI's adverse effects on labor. Comparative experiences highlighted in this essay reveal that both policies, like social insurance, training and education, and taxation, and institutions, like collective bargaining, can accelerate, offset, moderate, or intensify these effects.

Contemporary AI uses advanced computation to automate specific tasks at or above human cognitive capacity. Its development rests on advances in computing power and hardware, the proliferation of vast data sets, and evolving algorithms for analyzing and drawing statistical inferences and predictions.[2] Powered by machine learning (ML), recent AI breakthroughs are achieving human-comparable results across an expanding range of human tasks.[3] Despite remarkable advances, however, current AI applications remain narrow and task specific, with little ability to transfer "learning" from one problem to another. Narrow AI can displace humans in *low-level cognitive demand* tasks that are repetitive, data intensive, optimization-based, and asocial, but it cannot yet substitute for humans in most *high-level cognitive demand* tasks involving reasoning, real-world knowledge, judgment, and social interactions.

Narrow AI is also impacting human tasks both by adding intelligence to robots and production systems and by powering digital platforms that facilitate transactions between buyers and sellers. In a self-reinforcing cycle of data collection, analysis, and prediction, AI is driving the growth of digital platforms like Amazon for selling goods, Netflix for selling video services, and Uber and Upwork for selling labor services. Indeed, large tech platform companies in the United States and China, with massive amounts of data, extensive digital platforms, and relatively small employment compared with their revenues, currently account for about two-thirds of all business investment in AI.

"Artificial general intelligence" (AGI) has no clear definition, no clear time frame, is not yet a matter of business and policy concerns, and is therefore not our focus.[4] Instead, we concentrate on how narrow but rapidly evolving AI applications are likely to affect labor and livelihoods over the coming decade. Since even narrow AI applications are not yet widespread, our analysis by necessity draws on evidence about the impacts of other forms of automation on labor in the advanced industrial countries during the last three decades.

Predictions about the future effects of AI are replete with uncertainties both about the pace and scope of future scientific breakthroughs and about the pace and breadth of AI deployment. *Scientific advances determine whether a human task is technically automatable, but they do not determine whether it will be automated. That depends on deployment decisions.* In market economies, businesses make most of these decisions shaped by their strategies and the market, institutional,

and policy environments in which they operate. High taxes on labor relative to taxes on machinery and software, for example, have been a significant driver of business investments in automation technologies. Nonetheless, their deployment has been gradual because of substantial lags in the development of organizational capacities required for their effective utilization.[5]

Both historical evidence and economic logic indicate that on its current trajectory, AI will continue, intensify, and accelerate automation's adverse effects on key labor market trends in the advanced economies. These effects include the polarization of employment and wages, slow wage growth for middle- and low-skill workers, a significant premium in the wages of highly educated workers, a decoupling of wage growth from productivity growth, a decline in labor's share of value added, and growing income inequality.

Automation is not the only force behind these trends. Globalization, outsourcing, the decline in unionization and collective bargaining coverage, and the growing monopsony or "wage-setting" power of businesses are also significant factors.[6] These factors in turn have been enabled or reinforced by automation. Globalization and outsourcing, for example, have been turbocharged by robots and digitization in logistics, transportation, and communication.[7] By enabling the outsourcing of routine jobs to low-wage locations, networked technologies and automation, underpinned by the digital revolution, have propelled globalization, decreasing employment and constraining wage growth in manufacturing and tradable services in the advanced economies.

Yet while these factors have been at play in all of the advanced countries, there have been important differences among them in the consequences for labor. The varied outcomes have resulted in part from differences in policies, institutions, and societal norms of fairness. Strong competitive economies like Germany, Sweden, Canada, and Denmark have experienced the same technological and globalization forces as the United States, but their workers have fared much better.[8]

Examining *tasks* within occupations is a widely used optic to understand the impact of automation on labor, and we organize our analysis with this approach. Occupations encompass numerous tasks, only some of which are automatable or AI-susceptible.

Like other automation tools, AI impacts human tasks through three broad effects: the *displacement* effect, or the decrease in demand for labor in tasks that are automated; the *productivity* effect, or the increase in the demand for labor in nonautomated tasks; and the *reinstatement* effect, or the creation of new tasks for labor. Over time, but at a highly uncertain pace, automation's displacement effects are offset to some extent by both its productivity and reinstatement effects.[9]

The displacement effects can be immediate, significant, and palpable, and are themselves negative for employment and labor's share in value added. In contrast, the productivity and reinstatement benefits can take years, even decades, to mate-

rialize with significant frictional and structural unemployment, wage losses, and growing inequality along the way. In the long run – that ill-defined concept frequently used by economists – automation, productivity growth, and rising employment and wages move together.

But automation always involves disruption and, with it, winners and losers. Trade-offs can and do persist for "shorter" time horizons relevant to businesses, workers, citizens, and political leaders, and they have economic, social, and political consequences. History is replete with evidence that the social and political costs of labor market disruptions triggered by technological change can be significant.[10] And while displacement effects may hit particular locations or regions, productivity and reinstatement benefits can occur elsewhere: the costs often fall in one place and the benefits in another, complicating politics and policy.[11]

During the last thirty years, there is evidence that while automation's displacement effects have accelerated and intensified, its productivity and reinstatement effects have been slower to materialize and smaller than expected.[12] *The social and economic dislocations have grown, while the offsetting benefits have not been as robust or rapid as anticipated and have not been broadly shared.* In the United States, for example, despite growing automation and computerization of work, productivity growth slowed by nearly half to an annual rate of 1.5 percent during the last half century, and industries that led in the use of new information and communication technologies did not perform better in terms of total factor productivity, output, or employment growth. Nor is the United States alone: other advanced industrial economies have also experienced slowing productivity growth, the causes of which remain uncertain and robustly debated.[13]

Much of the automation during the last three decades is often called "routine-biased technological change," or RBTC, because it has substituted for humans in routine physical and increasingly routine cognitive tasks while increasing the demand for humans in nonroutine tasks. Both routine manual and routine cognitive occupations as a share of employment have fallen over the last thirty years. RBTC has been particularly important in automating tasks in structured, predictable environments like automobile production in factories and bookkeeping in offices.

We characterize AI as "RBTC on steroids" for two reasons. First, AI is adding intelligence to robots and other forms of automation that substitute for humans in routine and increasingly nonroutine physical tasks – think assembly-line production and warehousing. Second, AI is substituting for humans in a widening array of both routine and increasingly nonroutine cognitive tasks.[14] Cognitive tasks that are currently technically feasible for AI tend to be routine, data-intensive, and asocial (such as customer support, basic office support, and insurance underwriting). Physical tasks that are technically feasible for AI also tend to be routine, data-intensive, optimization-based, and asocial, and require limited dexterity and a structured environment (like assembly-line inspection or fruit harvesting). Most

high-level cognitive demand tasks in which inputs and outputs are not easily quantifiable with data, and which require both social interaction and cross-domain thinking, complex strategy, or creativity (such as the work of business and health professionals, teachers, and artists) are not directly in the crosshairs of current AI.[15]

If, as we conjecture, AI is RBTC on steroids, then its future effects on labor will be similar to the effects on labor from other forms of RBTC automation during the past thirty years. The first of these effects is the "polarization" of employment and, to a lesser extent, of wages. Many of the occupations hollowed out by RBTC over the previous three decades have been in manufacturing, which provided good jobs for millions of middle-skill, middle-wage workers. Polarization is reflected in a decline in the share of middle-skill occupations in total employment and increases in the employment shares of both low-skill and high-skill occupations, with the largest gains in the latter.[16] Although RBTC has been polarizing, it has been "upgrading" or "upskilling" in the sense that the decline in middle-skill occupations has been largely offset by an increase in high-skill occupations as shares of total employment.[17]

Polarization, in turn, has contributed to widening wage gaps among workers, with slow, stagnant, or even negative wage growth for workers whose occupations have been displaced by automation, and wage growth for those whose occupations have been enhanced by productivity gains or by the creation of new tasks. Earnings inequality has grown across the advanced industrial economies, largely driven by the rising pay gap or education premium between workers with a college-level education or rigorous training (like apprenticeships in Germany) whose skills have been complemented by RBTC and those with lower levels of education or training whose skills have been displaced.[18]

As a result of its sizable displacement and polarization effects, RBTC automation has also been a factor behind the decoupling of wage growth from productivity growth.[19] In theory, in competitive labor markets, wage growth should be commensurate with productivity growth in the long run, but productivity growth has outpaced both average and median wage growth over the past three decades. As noted earlier, the long run can be very long indeed, and there are large and lengthy aberrations along the path to getting there. Moreover, labor markets are usually not competitive, as narrowly defined by economists, and the sharing of productivity gains with workers depends not only on market forces but on the relative power of workers and employers. Relative power in turn is often reflected in tax and social policies and in institutions like corporate governance rules that favor owners over workers.

The decoupling of wage and productivity growth has contributed to a decline in labor's share of national income.[20] Indeed, automation has been a major driver of the decline in labor share most acute in manufacturing, and within manufacturing, most acute in industries undergoing rapid automation. In addition,

a declining labor share of national income has been mirrored in a rising capital share, further increasing income inequality, since capital returns are concentrated at the upper end of the income distribution.

The slow growth of pretax market incomes for the bottom 95 percent of wage earners has been the main driver of increasing income inequality in the advanced market economies over the past half century, and automation has played a major role.[21] The United States has been an outlier: no other advanced industrial economy has experienced an equally large rise in income inequality or equally severe wage stagnation for rank-and-file workers. Both eroding union coverage and a declining real minimum wage have been important factors behind the comparatively large gap between productivity growth and median wage growth, the comparatively large earnings inequalities by education, and the significant real wage decline for low-educated male workers in the United States. In contrast, in Germany, another large competitive market economy experiencing the same RBTC automation and globalization forces, broad collective bargaining rights, works councils, a generous social insurance system, a robust training system, and a national minimum wage have mitigated the adverse effects of automation on the supply of good jobs and have fostered more inclusive growth.[22]

Overall, RBTC automation has contributed to rising income inequality through a number of channels. It has resulted in stagnant or falling real wages for middle- and low-skill workers, favoring wages of high-skill workers complemented by automation; it has driven a large and persistent gap between wage growth and productivity growth; it has reduced labor's share and increased capital's share in value added; and it has produced "winner-take-all" income gains for superstar innovators and superstar firms with significant product market and monopsony power, contributing to rising income inequality both among them and between them and their workers.[23]

All of these factors are "market" explanations of wage stagnation and income inequality that reflect changes in the demand for different types of labor and capital resulting from RBTC automation. We are concerned that these market factors are likely to persist and indeed may strengthen as RBTC on steroids reduces the demand for labor with low and middle skills (and wages) performing both physical and cognitive routine tasks while increasing the demand for labor with skills required for nonroutine tasks of both types.

At the same time, we recognize that AI is likely to make human work more productive in some existing tasks and to create new tasks requiring human skills that cannot be replaced by AI capabilities. Uniquely human skills not susceptible to AI currently include social/interpersonal skills (teachers, care and health care workers, physical therapists, and hairdressers); physical skills in unpredictable environments (construction workers and plumbers); and general intelligence skills required for nonroutine tasks and problem-solving (management and artists).

For many occupations, the future of work is likely to involve growing interdependence between human skills and AI skills: for example, between the interpersonal skills of doctors and teachers with the complementary AI skills of data analysis, diagnostics, and prediction. Such complementary or partnership occupations in turn are likely to require high-level education and/or technical training for the human partners. Overall, such occupational changes are likely to fuel wage and income inequality between those workers whose skills are displaced by AI and those whose skills are complemented. A key but unanswered question is how the rewards from work will be shared between humans and their partner intelligent tools, between labor and the owners and creators of these tools.

So far, we have focused on how AI is affecting labor demand through the automation of tasks and occupations. Now we broaden our focus to consider how AI is affecting labor through enabling digital platforms that are creating new tasks and new forms of organizing work.[24] We believe that digital platforms, the use of which surged during the COVID-19 pandemic, will expand rapidly. To predict AI's future effects on labor, therefore, it is necessary to look through the lens of digital platforms. AI is enabling three types of digital platforms.

- *Platforms for selling goods* (such as Amazon and Netflix) recast what tasks are performed by humans and where. Accelerated by COVID-propelled changes in business practices to reduce workplace density and provide contactless service to customers, transactions continue to move from in-person, brick-and-mortar retail to e-commerce and digital platforms, with tasks shifting from shop floors to warehouse operations and long- and short-haul delivery and transportation.
- *Platforms for labor services* (such as Upwork, Lyft, and TaskRabbit), which utilize algorithms and real-time data to match workers with tasks, are having a growing impact on labor across industries. These platforms cover a wide range of tasks spanning nonroutine cognitive work like accounting and software work, nonroutine physical and technical work like electrical and plumbing services, and routine personal services like transportation and care.[25] Workers typically are matched with tasks for multiple clients, usually on a temporary project basis. Such work is often referred to as "gig work." Gig workers, including the digital assembly-line "ghost workers" who provide much of the human intelligence behind AI software, are part of the "on-demand gig economy."[26] And in response to COVID, new work-related platforms from Google to Zoom are expanding to facilitate remote or hybrid work for cognitive tasks.
- *Platforms for renting out assets* (such as Airbnb and BlaBlaCar) also offer new labor and income opportunities, even while they alter the character of work and the skills required for tasks.

Platform-mediated work is growing rapidly as a share of nonstandard employment arrangements (including independent contractors, temporary and on-call workers, and part-time workers) that already account for 25–31 percent of the working age populations in the advanced economies.[27] More than half of those participating in such arrangements use income from them to supplement their income from other sources. The platform-mediated gig portion of nonstandard employment arrangements is still small, accounting for an estimated 1–3 percent of total employment, but it is expanding quickly.[28]

Gig workers lack the legal and social protections provided in standard employment contracts, resulting in precarious jobs with low and unstable incomes, limited access to social insurance, minimal training and career development opportunities, exposure to health and safety risks, and low to zero collective bargaining rights.

As AI-enabled platforms transform relationships between employers and workers, new ways to finance and deliver social and legal protections are required to make gig and other platform jobs "good jobs." When COVID sharply reduced the demand for gig workers, most of the advanced economies added temporary measures, like pandemic unemployment benefits in the United States, to compensate workers for lost income.

Pre-COVID, many governments in Europe and a few U.S. states were already working on permanent measures to protect or empower gig workers. The United Kingdom, for example, added a new "worker" category, distinct from both the traditional employee category and the self-employed category, to its labor law. Some European countries are exploring extending social protections usually associated with standard employment contracts – such as unemployment and disability insurance, health coverage, and parental leave – to gig workers on labor services platforms.[29] Such benefits could be provided and financed through new "portable benefits programs," allowing workers to accumulate benefits on a prorated basis for time worked for different employers.[30]

Looking to the future, two forces will shape the demand for human labor in different tasks and occupations: the demand for goods and services that people want and the capabilities of intelligent tools and systems, empowered by AI, to produce and deliver them. Based on these two forces, over the next decade, we anticipate shifts in the composition of employment in the advanced industrial economies from occupations like office support, production, and warehousing that consist of many routine tasks to occupations in health care, education, technology, and the arts that encompass many nonroutine tasks.[31] The upskilling of employment is likely to continue with job growth concentrated in high-wage occupations and job declines in low- and middle-wage ones, further polarizing the labor market and fueling wage inequality. And the displacement and transition costs for workers who lose their jobs to AI and automation and who require different skills for new jobs are likely to be substantial, raising the question of who should bear these costs.[32]

These predicted shifts in occupations and their labor market effects are likely in both manufacturing and services that together account for more than 90 percent of employment. Manufacturing has been the locus of the hollowing out of "good" middle-skill, middle-wage jobs during the past thirty years, driven by robots, RBTC, and globalization. While manufacturing employment has fallen as a share of total employment, manufacturing output has not fallen nearly as sharply as a share of GDP. There have been significant productivity gains from automation, but they have not been broadly shared. A disproportionate share has gone to capital, not to workers, as evidenced by both the rising gap between productivity growth and wage growth and the fall in labor's share of value added. Moreover, the declines in manufacturing employment and wages have fallen hardest on workers in the lower half of the earnings distribution, on workers with less than a college degree, and on locations or regions in which manufacturing was a significant share of economic activity.

Similar disparities in the distribution of both displacement costs and productivity benefits are likely as AI drives further automation of manufacturing. Overall, the hollowing out of manufacturing jobs is likely to continue but also to be smaller than what occurred during the last thirty years. A new wave of AI-powered automation with increasingly programmable, semi-dexterous, and interconnected machines will optimize production systems. The resulting changes are likely to affect manufacturing employment by optimizing tasks that have already been automated and by creating new complementary tasks with required new skills for workers to operate new smarter systems. The pace at which manufacturing tasks are automated will depend not only on evolving AI capabilities but also on the improved dexterity of robots and production systems. Overall, AI is not likely to add significant risk of additional job displacement to "shop-floor" manufacturing workers, but it is likely to displace workers doing routine cognitive tasks in back offices.[33]

Based on both rising incomes and changing demographics, the demand for services will remain robust in the advanced industrial societies. Indeed, services already account for most (more than 80 percent) employment and almost all employment growth during the last several decades. Service occupations run the gamut from highly paid health and business professionals to middle-wage educators to low-wage retail clerks and hospitality workers. Given the diverse character of the service sector, we highlight briefly some of AI's implications in two large service industries: retail and health.

Artificial intelligence is transforming the *retail industry* across its value chain. On the demand side, businesses are shifting from traditional in-store channels to e-commerce channels, especially digital platforms, to anticipate demand and personalize the customer experience. On the supply side, AI is being applied to improve inventory forecasts, optimize merchandising and

product assortment, and automate warehousing and store operations. Overall employment in retail is likely to continue to decline, but the demand for humans in routine and nonroutine cognitive tasks in such areas as customer service, management, and technology deployment and maintenance is likely to increase. In contrast, routine manual jobs such as cashiers, drivers, packers, and shelf stockers are projected to decline, reducing low- to middle-wage job opportunities for workers with only a secondary education. In both manufacturing and services, the pace of change in AI-enabled drones and autonomous vehicles will impact the pace at which human tasks and wages in short-haul and long-haul transportation, two major middle-wage occupations, are affected.

In the *health care sector*, job growth is likely to remain strong. Indeed, both pre and post COVID, the health sector has topped the list of projected job growth in the advanced economies. Health care jobs cover a broad range of skills and incomes, from low-skill, low-wage jobs like orderlies and home care assistants through middle-skill, middle-wage jobs like lab technicians and paramedics to high-skill jobs like nurses, dentists, radiology technologists, and physicians. All of these job categories are projected to grow to keep pace with rising demand for health care services.

Within health care, AI is likely to complement the demand for high-wage workers performing nonroutine tasks requiring specialized skills and education while substituting for workers performing routine tasks. In particular, AI applications are likely to substitute for humans in data-dependent cognitive tasks in administrative and office support activities and patient relationship management while increasing the demand for humans in work performed by health professionals like nurses, doctors, physical therapists, and dentists whose responsibilities require high-level cognitive and/or highly skilled physical and social interaction tasks. The automation of administrative and data collection tasks, further enabled by telemedicine platforms, could be transformative for nurses who spend on average a quarter of their time on such duties, empowering them to use AI-informed results to offer more real-time health advice, diagnosis, and treatment.[34]

Many health care occupations are likely to require collaboration between humans with the requisite social skills and intelligent tools with the requisite data capabilities to deliver state-of-the art personalized services at scale. The scope for collaboration between humans and AI in health care is already apparent in the utilization of AI-enabled robots to address the interrelated demographic challenges of aging and shrinking populations. Japan, for example, is leading the way in robot use in tasks in nursing homes and hospitals, both to fill gaps in the supply of human labor available for these tasks and to complement the humans required to do them.

Throughout this essay, we have focused on the effects of AI and automation on the composition of *demand* for human labor in tasks, occupations, and jobs. Yet, as the example of Japan's adoption of robots in health care illus-

trates, employment, wages, and good jobs depend not only on the demand for human labor but also on its supply.[35] All of the industrial economies face a slowdown in the growth of their working age populations, albeit to differing degrees, and this is likely to result in shortages and upward pressure on wages both in occupations and jobs that are not currently susceptible to substitution by AI and in those that are complemented or enabled by it. As labor markets recover from COVID, there is already concern in the United States and in several European countries about future shortages of workers with the skills and education required to meet demand in growing sectors like health care and software engineering. Such shortages in turn are likely to accelerate innovation, investment, and deployment of AI-enabled automation technologies to substitute for human labor.

AI and the intelligent tools and systems it enables will automate many routine tasks, change existing tasks, and create new tasks for humans, often involving new forms of human and machine collaboration and new forms of work organization. There will be – indeed there already are – both winners and losers in this process of ongoing structural change. It is not sufficient to assert that as AI technologies transform work, there will ultimately be broad economic gains that are widely shared. That is not a technologically determined outcome but rather a societal choice. To foster both economic growth and the social and economic equity on which their prosperity and political stability depend, the advanced market economies must develop policies to share the disruption costs and productivity benefits of AI broadly, consistent with societal norms of fairness.

The availability and accessibility of good jobs should be core policy goals, yet achieving them is not trivial. To maximize the odds for success and to transform all jobs into good jobs, three broad types of policy interventions are warranted. First are lifelong education and training policies to equip workers with the skills they need for *access* to good jobs, along with active labor market policies to help them *transition* to these jobs. Second is the extension of social benefits and legal protections to cover workers in all businesses, including platform businesses. And third is a combination of income-support policies, including minimum wages, tax credits for work, and basic income supplements, to raise the after-tax earnings of workers who remain in low-wage jobs – including many routine service jobs in leisure and hospitality, health care, and childcare, many of them held by women and low-educated workers – to livelihood levels.[36]

Finally, it is important to emphasize that the effects of AI on work are not technologically determined but depend on the incentives of both those leading AI research and innovation and those investing in AI deployment. The prevailing narrative behind AI innovation and deployment in the business and research communities, a narrative particularly pronounced in the United States, where decisions reflect shareholder interests and workers have limited voice in business decisions, focuses on AI's ability to outperform humans, not on the creation of good jobs.

This narrative has been fostered by tax policies that raise the cost of labor and reduce the cost of capital, encouraging businesses to focus on automation technologies that reduce employment and cut labor costs without offsetting labor productivity growth. R&D tax incentives and other forms of government support for research in labor-saving technologies have reinforced the narrative, but well-designed policies could change it.[37]

Ultimately, how the economic benefits of intelligent machines and tools are realized and shared depend not on their technological design but on the design of intelligent policies needed for an inclusive AI era.[38]

AUTHORS' NOTE

The authors are grateful to the Ewing Marion Kauffman Foundation, the German Federal Ministry of Labour and Social Affairs (BMAS), and the American Academy in Berlin for their generous support for this work. Our research was a product of collaboration with the Hertie School, Humboldt Institute for Internet and Society, Weizenbaum Institute, and WZB Berlin Social Science Center. We thank Dafna Bearson and Camille Carlton for their outstanding research assistance.

ABOUT THE AUTHORS

Laura D. Tyson, a Fellow of the American Academy since 2002, is Distinguished Professor of the Graduate School at the Haas School of Business and Founder and Co-Faculty Director of the Sustainable and Impact Finance Initiative at the University of California, Berkeley. She also chairs the Board of Trustees at UC Berkeley's Blum Center for Developing Economies. She served in the Clinton administration as the Chair of the Council of Economic Advisers and as Director of the National Economic Council. She is the author of many publications, including *Who's Bashing Whom: Trade Conflicts in High-Technology Industries* (1992).

John Zysman is Professor Emeritus in the Department of Political Science at the University of California, Berkeley. His publications include *The Highest Stakes: The Economic Foundations of the Next Security System* (1992), *Manufacturing Matters: The Myth of the Post-Industrial Economy* (1987), and *Governments, Markets, and Growth: Finance and the Politics of Industrial Change* (1983).

ENDNOTES

[1] For a comprehensive analysis of automation's effects, see David Autor, David Mindell, and Elisabeth Reynolds, *The Work of the Future: Building Better Jobs in an Age of Intelligent Machines* (Cambridge, Mass.: MIT Press, 2020); and James Manyika, Susan Lund, Michael Chui, et al., *Jobs Lost, Jobs Gained: Workforce Transitions in a Time of Automation* (New York: McKinsey Global Institute, 2017).

[2] Thomas W. Malone, Daniela Rus, and Robert Laubacher, "Artificial Intelligence and the Future of Work" (Cambridge, Mass.: MIT Future of Work, 2020); Michael I. Jordan, "Artificial Intelligence–The Revolution Hasn't Happened Yet," *Harvard Data Science Review* 1 (1) (2019); Gary Marcus, "The Next Decade in AI: Four Steps Towards Robust Artificial Intelligence," arXiv (2020), https://arxiv.org/abs/2002.06177; John Zysman and Mark Nitzberg, *Governing AI: Understanding the Limits, Possibilities, and Risks of AI in an Era of Intelligent Tools and Systems* (Washington, D.C.: Wilson Center, 2020); Mark Nitzberg and John Zysman, "Algorithms, Data, and Platforms: The Diverse Challenges of Governing AI," BRIE Working Paper 2021-1 (Berkeley: Berkeley Roundtable on the International Economy, University of California, Berkeley, 2021); and Kevin Roose, *Futureproof: 9 Rules for Humans in the Age of Automation* (New York: Random House, 2021).

[3] Machine learning is the primary computer science breakthrough enabling contemporary AI. ML is a form of context-dependent statistical inference, in which algorithms are trained on vast amounts of data and improve ("learn") automatically through training. Deep Learning is a machine learning method that adjusts "weights" in multiple layers of artificial neural networks (ANNs) based on training data, and has powered the recent breakthrough, human-comparable AI results.

[4] While not dismissive of the possibility of AGI, many scholars agree that the focus of policy and regulation should be on the impacts of narrow AI. See, for example, Gary Marcus and Ernest Davis, *Rebooting AI: Building Artificial Intelligence We Can Trust* (New York: Pantheon Books, 2019); and Erik Brynjolfsson, Tom Mitchell, and Daniel Rock, "What Can Machines Learn, and What Does It Mean for Occupations and the Economy?" *AEA Papers and Proceedings* 108 (2018): 43–47.

[5] Erik Brynjolfsson, Daniel Rock, and Chad Syverson, "Understanding and Addressing the Modern Productivity Paradox" (Cambridge, Mass.: MIT Future of Work, 2020).

[6] Maarten Goos, Alan Manning, and Anna Salomons, "Explaining Job Polarization: Routine-Biased Technological Change and Offshoring," *American Economic Review* 104 (2) (2014): 2509–2526.

[7] Laura D. Tyson and Michael Spence, "Exploring the Effects of Technology on Income and Wealth Inequality," in *After Piketty: The Agenda for Economics and Inequality*, ed. Heather Boushey, J. Bradford DeLong, and Marshall Steinbaum (Cambridge, Mass.: Harvard University Press, 2017), 170–208.

[8] Anke Hassel and Bruno Palier, eds., *Growth and Welfare in Advanced Capitalist Economies: How Have Growth Regimes Evolved?* (Oxford: Oxford University Press, 2021); Autor et al., *The Work of the Future*; and Christian Dustmann, "Trade, Labor Markets, and the China Shock: What Can be Learned from the German Experience?" in *Combating Inequality: Rethinking Government's Role*, ed. Olivier Blanchard and Dani Rodrik (Cambridge, Mass.: MIT Press, 2021), 117–124.

[9] Daron Acemoglu and Pascual Restrepo, "Automation and New Tasks: How Technology Displaces and Reinstates Labor," *Journal of Economic Perspectives* 33 (2) (2019): 3–30. The

following lectures summarize the conclusions, with supporting data, on ongoing research by Acemoglu and Restrepo on automation, AI, and the effects on labor. Daron Acemoglu, "Tasks, Automation and Labor Market," presentation and lecture at Gorman Conference & Lectures 2020, virtual event, October 12, 2020; and Daron Acemoglu, "New Tasks, Good Automation and Bad Automation: Implications for the Future of Work," presentation and lecture at Gorman Conference & Lectures 2020, virtual event, October 13, 2020.

[10] This is well documented in Carl Benedikt Frey, *The Technology Trap: Capital, Labor, and Power in the Age of Automation* (Princeton, N.J.: Princeton University Press, 2019); and Barry Eichengreen, *The Populist Temptation: Economic Grievance and Political Reaction in the Modern Era* (New York: Oxford University Press, 2020).

[11] For evidence on the regional disparities of automation's effects in the United States, see Susan Lund, James Manyika, Liz Hilton Segel, et al., *The Future of Work in America: People and Places, Today and Tomorrow* (New York: McKinsey Global Institute, 2019). See also Mark Muro, *Countering the Geographical Impacts of Automation: Computers, AI, and Place Disparities* (Washington, D.C.: Brookings Institution, 2019).

[12] Daron Acemoglu, "Could We and Should We Reverse (Excessive) Automation?" in *Combating Inequality: Rethinking Government's Role*, ed. Olivier Blanchard and Dani Rodrik (Cambridge, Mass.: MIT Press, 2021), 163–170.

[13] Erik Brynjolfsson, Seth G. Benzell, and Daniel Rock, "How to Solve the Puzzle of Productivity Growth," Tech Stream, Brookings Institution, May 21, 2021.

[14] Daron Acemoglu, David Autor, Jonathon Hazell, and Pascual Restrepo, "AI and Jobs: Evidence from U.S. Vacancies," VoxEU, March 3, 2021, https://voxeu.org/article/ai-and-jobs-evidence-us-vacancies; Michael Webb, "The Impact of Artificial Intelligence on the Labor Market," working paper (2019); Erik Brynjolfsson, Tom Mitchell, and Daniel Rock, "What Can Machines Learn, and What Does It Mean for Occupations and the Economy?" *AEA Papers and Proceedings* 108 (2018): 43–47; and Edward W. Felten, Manav Raj, and Robert Seamans, "A Method to Link Advances in Artificial Intelligence to Occupational Abilities," *AEA Papers and Proceedings* 108 (2018): 54–57.

[15] Kai-Fu Lee, *AI Superpowers: China, Silicon Valley, and the New World Order* (Boston: Houghton Mifflin Harcourt, 2018); Kai-Fu Lee and Chen Qiufan, *AI 2041: Ten Visions for Our Future* (New York: Currency, 2021); Autor et al., *The Work of the Future*; and Ajay Agrawal, Joshua Gans, and Avi Goldfarb, eds., *The Economics of Artificial Intelligence: An Agenda* (Chicago: University of Chicago Press, 2019).

[16] The share of middle-skill occupations in total employment declined (11 percent), high-skill occupations grew (9 percent), and low-skill occupations grew (3 percent) between the mid-1990s to the late 2010s across OECD countries. See Organisation for Economic Co-operation and Development, *OECD Employment Outlook 2020: Worker Security and the COVID-19 Crisis* (Paris: OECD Publishing, 2020), Figure 4.1, https://doi.org/10.1787/1686c758-en.

[17] Georgios Petropoulos, "Occupational Change, Artificial Intelligence and the Geography of EU Labor Markets," Bruegel Working Paper 03 (Brussels: Bruegel, 2020).

[18] Florian Hoffmann, David S. Lee, and Thomas Lemieux, "Growing Income Inequality in the United States and Other Advanced Economies," *Journal of Economic Perspectives* 344 (4) (2020): 52–78; and Autor et al., *The Work of the Future*.

[19] David Autor and Anna Salomons, "Is Automation Labor-Displacing? Productivity Growth, Employment, and the Labor Share," Working Paper 24871 (Cambridge, Mass.: National Bureau of Economic Research, 2018); Mathilde Pak and Cyrille Schwellnus, "Labour Share Developments Over the Past Two Decades: The Role of Public Policies," OECD Economics Department Working Papers No. 1541 (Paris: OECD Publishing, 2019); Cyrille Schwellnus, Andreas Kappeler, and Pierre-Alain Pionnie, "The Decoupling of Median Wages from Productivity in OECD Countries," *International Productivity Monitor* 32 (2017); and Organisation for Economic Co-operation and Development, "Decoupling of Wages from Productivity: What Implications for Public Policies?" *OECD Economic Outlook* 2 (2018).

[20] On factors that contribute to the declining labor share, see Bennet Berger and Guntram B. Wolff, "The Global Decline in the Labour Income Share: Is Capital the Answer to Germany's Current Account Surplus?" (Brussels: Bruegel, 2017); James Manyika, Jan Mischke, Jacques Bughin, et al., "A New Look at the Declining Labor Share of Income in the United States" (New York: McKinsey Global Institute, 2019); and Caterina Astarita and Gaetano D'Adamo, "Inequality and Structural Reforms: Methodological Concerns and Lessons from Policy" (Brussels: Directorate-General for Economic and Financial Affairs, European Commission, 2017), https://ideas.repec.org/p/euf/dispap/071.html.

[21] Lucas Chancel, "Ten Facts about Inequality in Advanced Economies," in *Combating Inequality: Rethinking Government's Role*, ed. Olivier Blanchard and Dani Rodrik (Cambridge, Mass.: MIT Press, 2021), 3–30.

[22] Laura Tyson, "Automation and the Future of Work in Germany: A Summary of Research and Policy Recommendations?" Governing Work in the Digital Age Project Working Paper (Berlin: Hertie School, 2020), https://digitalage.berlin/wp-content/uploads/2021/12/2021_Tyson-WP1-Paper-FINAL-1.pdf; and Laura Tyson, "The Future of Work in Germany," *The Berlin Journal* 33 (2019).

[23] Tyson and Spence, "Exploring the Effects of Technology on Income and Wealth Inequality"; and Anton Korinek and Joseph E. Stiglitz, *Artificial Intelligence and Its Implications for Income Distribution and Unemployment* (Chicago: University of Chicago Press, 2019).

[24] Dafna Bearson, Martin Kenney, and John Zysman, "Measuring the Impacts of Labor in the Platform Economy: New Work Created, Old Work Reorganized, and Value Creation Reconfigured," *Industrial and Corporate Change* 30 (3) (2021), https://doi.org/10.1093/icc/dtaa046; and Martin Kenney, Dafna Bearson, and John Zysman, "The Platform Economy Matures: Measuring Pervasiveness and Exploring Power," *Socio-Economic Review* 19 (4) (2021), https://doi.org/10.1093/ser/mwab014.

[25] Jacques Bughin, Eric Hazan, Susan Lund, et al., *Skill Shift: Automation and the Future of the Workforce* (New York: McKinsey Global Institute, 2018).

[26] Mary Gray and Siddharth Suri, *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass* (Boston: Houghton Mifflin Harcourt, 2019); and Daisuke Wakabayashi, "Google's Shadow Workforce: Temps Who Outnumber Full Time Employees," *The New York Times*, May 28, 2019, https://www.nytimes.com/2019/05/28/technology/google-temp-workers.html?searchResultPosition=1.

[27] Jonathan I. Dingel and Brent Neiman, "How Many Jobs Can be Done at Home?" *Journal of Public Economics* 189 (2) (2020).

[28] Susan Lund, Anu Madgavkar, James Manyika, et al., *The Future of Work after COVID-19* (New York: McKinsey Global Institute, 2021).

[29] Organisation for Economic Co-operation and Development, *Employment Outlook* 2020; and International Labour Organization, "The Role of Digital Labour Platforms in Transforming the World of Work," in *World Employment and Social Outlook 2021* (Geneva: International Labour Organization, 2021), https://www.ilo.org/global/research/global-reports/weso/2021/lang--en/index.htm.

[30] Libby Reder, Shelly Steward, and Natalie Foster, *Designing Portable Benefits: A Resource Guide for Policymakers* (Washington, D.C.: Aspen Institute, 2019); and Steven Hill, *Raw Deal: How the "Uber Economy" and Runaway Capitalism Are Screwing American Workers* (New York: St. Martin's Press, 2015).

[31] Three changes in business and consumer practices in response to COVID-19 are likely to reinforce these changes in the composition of labor demand: the expansion of hybrid work for both routine and nonroutine cognitive tasks; the expansion of e-commerce, reducing demand for workers in low-wage routine tasks in retail/leisure/restaurants with partially offsetting increases in demand for workers doing warehousing, delivery, and other logistics tasks; and the acceleration of AI-enabled automation and digitalization of tasks to reduce workplace density and social interaction. See Lund et al., *The Future of Work after COVID-19*.

[32] For other recent studies predicting the future effects of AI on work, see Jason Furman and Robert Seamans, "AI and the Economy," NBER Working Paper 24689 (Cambridge, Mass.: National Bureau of Economic Research, 2018); Marguerita Lane and Anne Saint-Martin, "The Impact of Artificial Intelligence on the Labour Market: What Do We Know So Far?" OECD Social, Employment and Migration Working Papers No. 256 (Paris: OECD Publishing, 2021); and Michael Chui, *Artificial Intelligence – The Next Digital Frontier?* (New York: McKinsey and Company Global Institute, 2017), 3–6.

[33] Moreover, as AI enables sensors and data, it will create new manufacturing tasks for new manufactured products that are packaged and sold as "services for everything" bundles, combining physical goods with ongoing maintenance, upgrading, and replacement services.

[34] Thomas Davenport and Ravi Kalakota, "The Potential for Artificial Intelligence in Healthcare," *Future Healthcare Journal* 6 (2) (2019): 94; and Ari Bronsoler, Joseph Doyle, and John Van Reenen, "The Impact of New Technology on the Healthcare Workforce" (Cambridge, Mass.: MIT Work of the Future, 2020).

[35] Hal Varian, "Automation Versus Procreation (aka Bots Versus Tots)," VoxEU, March 30, 2020, https://voxeu.org/article/automation-versus-procreation-aka-bots-versus-tots.

[36] For an insightful framework distinguishing preproduction, production, and postproduction policies for inclusive growth, see Dani Rodrik and Stefanie Stantcheva, "Economic Inequality and Insecurity: Policies for an Inclusive Economy" (report prepared for commission chaired by Olivier Blanchard and Jean Tirole on major future economic challenges, Republic of France, 2021), 193–328.

[37] Daron Acemoglu and Pascual Restrepo, "The Wrong Kind of AI? Artificial Intelligence and the Future of Labour Demand," *Cambridge Journal of Regions, Economy and Society* 13 (1) (2019): 25–35.

[38] Erik Brynjolfsson and Andrew McAfee, *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies* (New York: W. W. Norton & Company, 2016); and Laura D'Andrea Tyson, "Intelligent Machines and Displaced Workers" (Emeryville, Calif.: Berkeley Research Group, 2014).

The Turing Trap: The Promise & Peril of Human-Like Artificial Intelligence

Erik Brynjolfsson

In 1950, Alan Turing proposed a test of whether a machine was intelligent: could a machine imitate a human so well that its answers to questions were indistinguishable from a human's? Ever since, creating intelligence that matches human intelligence has implicitly or explicitly been the goal of thousands of researchers, engineers, and entrepreneurs. The benefits of human-like artificial intelligence (HLAI) include soaring productivity, increased leisure, and perhaps most profoundly a better understanding of our own minds. But not all types of AI are human-like – in fact, many of the most powerful systems are very different from humans – and an excessive focus on developing and deploying HLAI can lead us into a trap. As machines become better substitutes for human labor, workers lose economic and political bargaining power and become increasingly dependent on those who control the technology. In contrast, when AI is focused on augmenting humans rather than mimicking them, humans retain the power to insist on a share of the value created. What is more, augmentation creates new capabilities and new products and services, ultimately generating far more value than merely human-like AI. While both types of AI can be enormously beneficial, there are currently excess incentives for automation rather than augmentation among technologists, business executives, and policy-makers.

Alan Turing was far from the first to imagine human-like machines.[1] According to legend, 3,500 years ago, Dædalus constructed humanoid statues that were so lifelike that they moved and spoke by themselves.[2] Nearly every culture has its own stories of human-like machines, from Yanshi's leather man described in the ancient Chinese *Liezi* text to the bronze Talus of the Argonautica and the towering clay *Mokkerkalfe* of Norse mythology. The word robot first appeared in Karel Čapek's influential play *Rossum's Universal Robots* and derives from the Czech word *robota*, meaning servitude or work. In fact, in the first drafts of his play, Čapek named them *labori* until his brother Josef suggested substituting the word robot.[3]

Of course, it is one thing to tell tales about humanoid machines. It is something else to create robots that do real work. For all our ancestors' inspiring stories, we are the first generation to build and deploy real robots in large numbers.[4]

© 2022 by Erik Brynjolfsson
Published under a Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) license
https://doi.org/10.1162/DAED_a_01915

Dozens of companies are working on robots as human-like, if not more so, as those described in the ancient texts. One might say that technology has advanced sufficiently to become indistinguishable from mythology.[5]

The breakthroughs in robotics depend not merely on more dexterous mechanical hands and legs, and more perceptive synthetic eyes and ears, but also on increasingly human-like artificial intelligence (HLAI). Powerful AI systems are crossing key thresholds: matching humans in a growing number of fundamental tasks such as image recognition and speech recognition, with applications from autonomous vehicles and medical diagnosis to inventory management and product recommendations.[6]

These breakthroughs are both fascinating and exhilarating. They also have profound economic implications. Just as earlier general-purpose technologies like the steam engine and electricity catalyzed a restructuring of the economy, our own economy is increasingly transformed by AI. A good case can be made that AI is the most general of all general-purpose technologies: after all, if we can solve the puzzle of intelligence, it would help solve many of the other problems in the world. And we are making remarkable progress. In the coming decade, machine intelligence will become increasingly powerful and pervasive. We can expect record wealth creation as a result.

Replicating human capabilities is valuable not only because of its practical potential for reducing the need for human labor, but also because it can help us build more robust and flexible forms of intelligence. Whereas domain-specific technologies can often make rapid progress on narrow tasks, they founder when unexpected problems or unusual circumstances arise. That is where human-like intelligence excels. In addition, HLAI could help us understand more about ourselves. We appreciate and comprehend the human mind better when we work to create an artificial one.

These are all important opportunities, but in this essay, I will focus on the ways that HLAI could lead to a realignment of economic and political power.

The distributive effects of AI depend on whether it is primarily used to augment human labor or automate it. When AI augments human capabilities, enabling people to do things they never could before, then humans and machines are complements. Complementarity implies that people remain indispensable for value creation and retain bargaining power in labor markets and in political decision-making. In contrast, when AI replicates and automates existing human capabilities, machines become better substitutes for human labor and workers lose economic and political bargaining power. Entrepreneurs and executives who have access to machines with capabilities that replicate those of humans for a given task can and often will replace humans in those tasks.

Automation increases productivity. Moreover, there are many tasks that are dangerous, dull, or dirty, and those are often the first to be automated. As more

tasks are automated, a fully automated economy could, in principle, be structured to redistribute the benefits from production widely, even to those people who are no longer strictly necessary for value creation. However, the beneficiaries would be in a weak bargaining position to prevent a change in the distribution that left them with little or nothing. Their incomes would depend on the decisions of those in control of the technology. This opens the door to increased concentration of wealth and power.

This highlights the promise and the peril of achieving HLAI : building machines designed to pass the Turing Test and other, more sophisticated metrics of human-like intelligence.[7] On the one hand, it is a path to unprecedented wealth, increased leisure, robust intelligence, and even a better understanding of ourselves. On the other hand, if HLAI leads machines to automate rather than augment human labor, it creates the risk of concentrating wealth and power. And with that concentration comes the peril of being trapped in an equilibrium in which those without power have no way to improve their outcomes, a situation I call the *Turing Trap*.

The grand challenge of the coming era will be to reap the unprecedented benefits of AI, including its human-like manifestations, while avoiding the Turing Trap. Succeeding in this task requires an understanding of how technological progress affects productivity and inequality, why the Turing Trap is so tempting to different groups, and a vision of how we can do better.

Artificial intelligence pioneer Nils Nilsson noted that "achieving real human-level AI would necessarily imply that most of the tasks that humans perform for pay could be automated."[8] In the same article, he called for a focused effort to create such machines, writing that "achieving human-level AI or 'strong AI' remains the ultimate goal for some researchers" and he contrasted this with "weak AI," which seeks to "build machines that help humans."[9] Not surprisingly, given these monikers, work toward "strong AI" attracted many of the best and brightest minds to the quest of – implicitly or explicitly – fully automating human labor, rather than assisting or augmenting it.

For the purposes of this essay, rather than strong versus weak AI, let us use the terms *automation* versus *augmentation*. In addition, I will use HLAI to mean human-*like* artificial intelligence, not human-*level* AI, because the latter mistakenly implies that intelligence falls on a single dimension, and perhaps even that humans are at the apex of that metric. In reality, intelligence is multidimensional : a 1970s pocket calculator surpasses the most intelligent human in some ways (such as for multiplication), as does a chimpanzee (short-term memory). At the same time, machines and animals are inferior to human intelligence on myriad other dimensions. The term "artificial general intelligence" (AGI) is often used as a synonym for HLAI. However, taken literally, it is the union of all types of intelligences,

able to solve types of problems that are solvable by any existing human, animal, or machine. That suggests that AGI is not human-like.

The good news is that both automation and augmentation can boost labor productivity: that is, the ratio of value-added output to labor-hours worked. As productivity increases, so do average incomes and living standards, as do our capabilities for addressing challenges from climate change and poverty to health care and longevity. Mathematically, if the human labor used for a given output declines toward zero, then labor productivity would grow to infinity.[10]

The bad news is that no economic law ensures everyone will share this growing pie. Although pioneering models of economic growth assumed that technological change was neutral,[11] in practice, technological change can disproportionately help or hurt some groups, even if it is beneficial on average.[12]

In particular, the way the benefits of technology are distributed depends to a great extent on how the technology is deployed and the economic rules and norms that govern the equilibrium allocation of goods, services, and incomes. When technologies automate human labor, they tend to reduce the marginal value of workers' contributions, and more of the gains go to the owners, entrepreneurs, inventors, and architects of the new systems. In contrast, when technologies augment human capabilities, more of the gains go to human workers.[13]

A common fallacy is to assume that all or most productivity-enhancing innovations belong in the first category: automation. However, the second category, augmentation, has been far more important throughout most of the past two centuries. One metric of this is the economic value of an hour of human labor. Its market price as measured by median wages has grown more than tenfold since 1820.[14] An entrepreneur is willing to pay much more for a worker whose capabilities are amplified by a bulldozer than one who can only work with a shovel, let alone with bare hands.

In many cases, not only wages but also employment grow with the introduction of new technologies. With the invention of the airplane, a new job category was born: pilots. With the invention of jet engines, pilot productivity (in passenger-miles per pilot-hour) grew immensely. Rather than reducing the number of employed pilots, the technology spurred demand for air travel so much that the number of pilots grew. Although this pattern is comforting, past performance does not guarantee future results. Modern technologies – and, more important, the ones under development – are different from those that were important in the past.[15]

In recent years, we have seen growing evidence that not only is the labor share of the economy declining, but even among workers, some groups are beginning to fall even further behind.[16] Over the past forty years, the numbers of millionaires and billionaires grew while the average real wages for Americans with only a high school education fell.[17] Though many phenomena contributed to this, including new patterns of global trade, changes in technology deployment are the single biggest explanation.

If capital in the form of AI can perform more tasks, those with unique assets, talents, or skills that are *not* easily replaced with technology stand to benefit disproportionately.[18] The result has been greater wealth concentration.[19]

Ultimately, a focus on more human-like AI can make technology a better substitute for the many nonsuperstar workers, driving down their market wages, even as it amplifies the market power of a few.[20] This has created a growing fear that AI and related advances will lead to a burgeoning class of unemployable or "zero marginal product" people.[21]

As noted above, both automation and augmentation can increase productivity and wealth. However, an unfettered market is likely to create socially excessive incentives for innovations that automate human labor and provide too weak incentives for technology that augments humans. The first fundamental welfare theorem of economics states that under a particular set of conditions, market prices lead to a *pareto optimal* outcome: that is, one where no one can be made better off without making someone else worse off. But we should not take too much comfort in that. The theorem does not hold when there are innovations that change the production possibilities set or externalities that affect people who are not part of the market.[22]

Both innovations and externalities are of central importance to the economic effects of AI, since AI is not only an innovation itself, but also one that triggers cascades of complementary innovations, from new products to new production systems.[23] Furthermore, the effects of AI, particularly on work, are rife with externalities. When a worker loses opportunities to earn labor income, the costs go beyond the newly unemployed to affect many others in their community and in the broader society. With fading opportunities often come the dark horses of alcoholism, crime, and opioid abuse. Recently, the United States has experienced the first decline in life expectancies in its recorded history, a result of increasing deaths from suicide, drug overdose, and alcoholism, what economists Anne Case and Angus Deaton call "deaths of despair."[24]

This spiral of marginalization can grow because concentration of economic power often begets concentration of political power. In the words attributed to Louis Brandeis: "We may have democracy, or we may have wealth concentrated in the hands of a few, but we can't have both." In contrast, when humans are indispensable to value creation, economic power will tend to be more decentralized. Historically, most economically valuable knowledge – what economist Simon Kuznets called "useful knowledge" – resided within human brains.[25] But no human brain can contain even a small fraction of the useful knowledge needed to run even a medium-sized business, let alone a whole industry or economy, so knowledge had to be distributed and decentralized.[26] The decentralization of useful knowledge, in turn, decentralizes economic and political power.

Unlike nonhuman assets such as property and machinery, much of a person's knowledge is inalienable, both in the practical sense that no one person can know everything that another person knows and in the legal sense that its ownership cannot be legally transferred.[27] In contrast, when knowledge becomes codified and digitized, it can be owned, transferred, and concentrated very easily. Thus, when knowledge shifts from humans to machines, it opens the possibility of concentration of power. When historians look back on the first two decades of the twenty-first century, they will note the striking growth in the digitization and codification of information and knowledge.[28] In parallel, machine learning models are becoming larger, with hundreds of billions of parameters, using more data and getting more accurate results.[29]

More formally, incomplete contracts theory shows how ownership of key assets provides bargaining power in relationships between economic agents (such as employers and employees, or business owners and subcontractors).[30] To the extent that a person controls an indispensable asset (like useful knowledge) needed to create and deliver a company's products and services, that person can command not only higher income but also a voice in decision-making. When useful knowledge is inalienably locked in human brains, so too is the power it confers. But when it is made alienable, it enables (though does not demand) greater concentration of decision-making and power.[31]

The risks of the Turing Trap are amplified because three groups of people – technologists, businesspeople, and policy-makers – each find automation alluring. Technologists have sought to replicate human intelligence for decades to address the recurring challenge of what computers could not do. The invention of computers and the birth of the term "electronic brain" were the latest fuel for the ongoing battle between technologists and humanist philosophers.[32] The philosophers posited a long list of ordinary and lofty human capacities that computers would never be able to do. No machine could play checkers, master chess, read printed words, recognize speech, translate between human languages, distinguish images, climb stairs, win at Jeopardy or Go, write poems, and so forth.

For professors, it is tempting to assign such projects to their graduate students. Devising challenges that are new, useful, and achievable can be as difficult as solving them. Rather than specify a task that neither humans nor machines have ever done before, why not ask the research team to design a machine that replicates an existing human capability? Unlike more ambitious goals, replication has an existence proof that such tasks are, in principle, feasible and useful.

While the appeal of human-like systems is clear, the paradoxical reality is that HLAI can be more difficult and less valuable than systems that achieve superhuman performance.

In 1988, robotics researcher Hans Moravec noted that "it is comparatively easy to make computers exhibit adult level performance on intelligence tests or playing checkers, and difficult or impossible to give them the skills of a one-year-old when it comes to perception and mobility."[33] But I would argue that in many domains, Moravec was not nearly ambitious enough. It is often comparatively easier for a machine to achieve *superhuman* performance in new domains than to match ordinary humans in the tasks they do regularly.

Humans have evolved over millions of years to be able to comfort a baby, navigate a cluttered forest, or pluck the ripest blueberry from a bush. These tasks are difficult if not impossible for current machines. But machines excel when it comes to seeing X-rays, etching millions of transistors on a fragment of silicon, or scanning billions of webpages to find the most relevant one. Imagine how feeble and limited our technology would be if past engineers set their sights on merely matching human-levels of perception, actuation, and cognition.

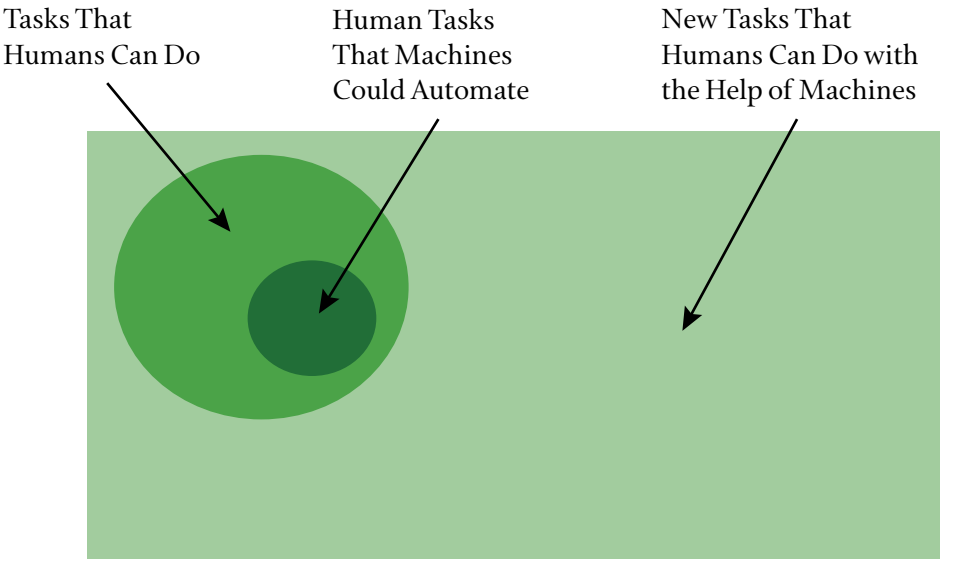
Augmenting humans with technology opens an endless frontier of new abilities and opportunities. The set of tasks that humans and machines can do together is undoubtedly much larger than those humans can do alone (Figure 1). Machines can perceive things that are imperceptible to humans, they can act on objects in ways that no human can, and, most intriguingly, they can comprehend things that are incomprehensible to the human brain. As Demis Hassabis, CEO of DeepMind, put it, the AI system "doesn't play like a human, and it doesn't play like a program. It plays in a third, almost alien, way . . . it's like chess from another dimension."[34] Computer scientist Jonathan Schaeffer explains the source of its superiority: "I'm absolutely convinced it's because it hasn't learned from humans."[35] More fundamentally, inventing tools that augment the process of invention itself promises to expand not only our collective abilities, but to accelerate the rate of expansion of those abilities.

What about businesspeople? They often find that substituting machinery for human labor is the low-hanging fruit of innovation. The simplest approach is to implement plug-and-play automation: swap in a piece of machinery for each task a human is currently doing. That mindset reduces the need for more radical changes to business processes.[36] Task-level automation reduces the need to understand subtle interdependencies and creates easy A-B tests, by focusing on a known task with easily measurable performance improvement.

Similarly, because labor costs are the biggest line item in almost every company's budget, automating jobs is a popular strategy for managers. Cutting costs – which can be an internally coordinated effort – is often easier than expanding markets. Moreover, many investors prefer "scalable" business models, which is often a synonym for a business that can grow without hiring and the complexities that entails.

But here again, when businesspeople focus on automation, they often set out to achieve a task that is both less ambitious and more difficult than it need be.

Figure 1
Opportunities for Augmenting Humans Are Far Greater than Opportunities to Automate Existing Tasks



To understand the limits of substitution-oriented automation, consider a thought experiment. Imagine that our old friend Dædalus had at his disposal an extremely talented team of engineers 3,500 years ago and built human-like machines that fully automated every work-related task that his fellow Greeks were doing.

- ✓ Herding sheep? Automated.
- ✓ Making clay pottery? Automated.
- ✓ Weaving tunics? Automated.
- ✓ Repairing horse-drawn carts? Automated.
- ✓ Incense and chanting for victims of disease? Automated.

The good news is that labor productivity would soar, freeing the ancient Greeks for a life of leisure. The bad news is that their living standards and health outcomes would come nowhere near matching ours. After all, there is only so much value one can get from clay pots and horse-drawn carts, even with unlimited quantities and zero prices.

In contrast, most of the value that our economy has created since ancient times comes from new goods and services that not even the kings of ancient empires had, not from cheaper versions of existing goods.[37] In turn, myriad new tasks are

required: fully 60 percent of people are now employed in occupations that did not exist in 1940.[38] In short, automating labor ultimately unlocks less value than augmenting it to create something new.

At the same time, automating a whole job is often brutally difficult. Every job involves multiple different tasks, including some that are extremely challenging to automate, even with the cleverest technologies. For example, AI may be able to read mammograms better than a human radiologist, but it is not very good at the other twenty-six tasks associated with the job, according to O-NET, such as comforting a concerned patient or coordinating on a care plan with other doctors.[39] My work with Tom Mitchell and Daniel Rock on the suitability for machine learning analyzed 950 distinct occupations. We found that machines could perform at least some tasks in most occupations, but zero in which machine learning could do 100 percent of the tasks.[40]

The same principle applies to the more complex production systems that involve multiple people working together.[41] To be successful, firms typically need to adopt a new technology as part of a system of mutually reinforcing organizational changes.[42] Consider another thought experiment: Imagine if Jeff Bezos had "automated" existing bookstores by simply replacing all the human cashiers with robot cashiers. That might have cut costs a bit, but the total impact would have been muted. Instead, Amazon reinvented the concept of a bookstore by combining humans and machines in a novel way. As a result, they offer vastly greater product selection, ratings, reviews, and advice, and enable 24/7 retail access from the comfort of customers' homes. The power of the technology was not in automating the work of humans in the existing retail bookstore concept but in reinventing and augmenting how customers find, assess, purchase, and receive books and, in turn, other retail goods.

Third, policy-makers have also often tilted the playing field toward automating human labor rather than augmenting it. For instance, the U.S. tax code currently encourages capital investment over investment in labor through effective tax rates that are much higher on labor than on plants and equipment.[43]

Consider a third thought experiment: Two potential ventures each use AI to create $1 billion of profits. If one of them achieves this by augmenting and employing a thousand workers, the firm will owe corporate and payroll taxes, while the employees will pay income taxes, payroll taxes, and other taxes. If the second business has no employees, the government may collect the same corporate taxes, but no payroll taxes and no taxes paid by workers. As a result, the second business model pays far less in total taxes.

This disparity is amplified because the tax code treats labor income more harshly than capital income. In 1986, top tax rates on capital income and labor income were equalized in the United States, but since then, successive changes have created a large disparity, with the 2021 top marginal federal tax rates on labor

income of 37 percent, while long capital gains have a variety of favorable rules, including a lower statutory tax rate of 20 percent, the deferral of taxes until capital gains are realized, and the "step-up basis" rule that resets capital gains to zero, wiping out the associated taxes, when assets are inherited.

The first rule of tax policy is simple: you tend to get less of whatever you tax. Thus, a tax code that treats income that uses labor less favorably than income derived from capital will favor automation over augmentation. Treating both business models equally would lead to more balanced incentives. In fact, given the positive externalities of more widely shared prosperity, a case could be made for treating wage income *more* favorably than capital income, for instance by expanding the earned income tax credit.[44] It is unlikely that any government official can define in advance exactly which technologies and innovations augment humans rather than merely substitute for them; indeed, most technologies have elements of each and the outcome depends a great deal on how they are deployed. Thus, rather than prescribe or proscribe specific technologies, a broad-based set of incentives can gently nudge technologists and managers toward augmentation on the margin, much as carbon taxes encourage myriad types of cleaner energy or research and development tax credits encourage greater investments in research.

Government policy in other areas could also do more to steer the economy clear of the Turing Trap. The growing use of AI, even if only for complementing workers, and the further reinvention of organizations around this new general-purpose technology imply a great need for worker training or retraining. In fact, for each dollar spent on machine learning technology, companies may need to spend nine dollars on intangible human capital.[45] However, education and training suffer from a serious externality issue: companies that incur the costs to train or retrain workers may reap only a fraction of the benefits of those investments, with the rest potentially going to other companies, including competitors, as these workers are free to bring their skills to their new employers. At the same time, workers are often cash- and credit-constrained, limiting their ability to invest in their own skills development.[46] This implies that government policy should directly provide education and training or provide incentives for corporate training that offset the externalities created by labor mobility.[47]

In sum, the risks of the Turing Trap are increased not by just one group in our society, but by the misaligned incentives of technologists, businesspeople, and policy-makers.

The future is not preordained. We control the extent to which AI either expands human opportunity through augmentation or replaces humans through automation. We can work on challenges that are easy for machines and hard for humans, rather than hard for machines and easy for humans. The first option offers the opportunity of growing and sharing the economic pie

by augmenting the workforce with tools and platforms. The second option risks dividing the economic pie among an ever-smaller number of people by creating automation that displaces ever-more types of workers.

While both approaches can and do contribute to productivity and progress, technologists, businesspeople, and policy-makers have each been putting a finger on the scales in favor of replacement. Moreover, the tendency of a greater concentration of technological and economic power to beget a greater concentration of political power risks trapping a powerless majority into an unhappy equilibrium : the Turing Trap.

The backlash against free trade offers a cautionary tale. Economists have long argued that free trade and globalization tend to grow the economic pie through the power of comparative advantage and specialization. They have also acknowledged that market forces alone do not ensure that every person in every country will come out ahead. So they proposed a grand bargain : maximize free trade to maximize wealth creation and then distribute the benefits broadly to compensate any injured occupations, industries, and regions. It has not worked as they had hoped. As the economic winners gained power, they reneged on the second part of the bargain, leaving many workers worse off than before.[48] The result helped fuel a populist backlash that led to import tariffs and other barriers to free trade. Economists wept.

Some of the same dynamics are already underway with AI. More and more Americans, and indeed workers around the world, believe that while the technology may be creating a new billionaire class, it is not working for them. The more technology is used to replace rather than augment labor, the worse the disparity may become, and the greater the resentments that feed destructive political instincts and actions. More fundamentally, the moral imperative of treating people as ends, and not merely as means, calls for everyone to share in the gains of automation.

The solution is not to slow down technology, but rather to eliminate or reverse the excess incentives for automation over augmentation. A good start would be to replace the Turing Test, and the mindset it embodies, with a new set of practical benchmarks that steer progress toward AI-powered systems that exceed anything that could be done by humans alone. In concert, we must build political and economic institutions that are robust in the face of the growing power of AI. We can reverse the growing tech backlash by creating the kind of prosperous society that inspires discovery, boosts living standards, and offers political inclusion for everyone. By redirecting our efforts, we can avoid the Turing Trap and create prosperity for the many, not just the few.

AUTHOR'S NOTE

The core ideas in this essay were inspired by a series of conversations with James Manyika and Andrew McAfee. I am grateful for valuable comments and suggestions on this work from Matt Beane, Seth Benzell, Avi Goldfarb, Katya Klinova, Alena Kykalova, Gary Marcus, Andrea Meyer, Dana Meyer, and numerous participants at seminars at the Stanford Digital Economy Lab and the University of Toronto Creative Destruction Lab, but they should not be held responsible for any errors or opinions in the essay.

ABOUT THE AUTHOR

Erik Brynjolfsson is the Jerry Yang and Akiko Yamazaki Professor and Senior Fellow at the Institute for Human-Centered AI and Director of the Digital Economy Lab at Stanford University. He is also the Ralph Landau Senior Fellow at the Institute for Economic Policy Research and Professor by Courtesy at the Graduate School of Business and Department of Economics at Stanford University; and a Research Associate at the National Bureau of Economic Research. He is the author or coauthor of seven books, including *Machine, Platform, Crowd: Harnessing Our Digital Future* (2017), *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies* (2014), and *Race against the Machine: How the Digital Revolution Is Accelerating Innovation, Driving Productivity, and Irreversibly Transforming Employment and the Economy* (2011) with Andrew McAfee, and *Wired for Innovation: How Information Technology Is Reshaping the Economy* (2009) with Adam Saunders.

ENDNOTES

[1] Alan Turing, "Computing Machinery and Intelligence," *Mind* 59 (236): 433–460, https://doi.org/10.1093/mind/LIX.236.433. An earlier articulation of this test comes from Descartes in *The Discourse*, in which he wrote,

> If there were machines which bore a resemblance to our bodies and imitated our actions as closely as possible for all practical purposes, we should still have two very certain means of recognizing that they were not real men. The first is that they could never use words, or put together signs, as we do in order to declare our thoughts to others. . . . Secondly, even though some machines might do some things as well as we do them, or perhaps even better, they would inevitably fail in others, which would reveal that they are acting not from understanding.

[2] Carolyn Price, "Plato, Opinions and the Statues of Daedalus," OpenLearn, updated June 19, 2019, https://www.open.edu/openlearn/history-the-arts/philosophy/plato-opinions-and-the-statues-daedalus; and Andrew Stewart, "The Archaic Period," Perseus Digital Library, http://www.perseus.tufts.edu/hopper/text?doc=Perseus:text:1999.04.0008:part=2:chapter=1&highlight=daedalus.

[3] "The Origin of the Word 'Robot,'" Science Friday, April 22, 2011, https://www.sciencefriday.com/segments/the-origin-of-the-word-robot/.

[4] Millions of people are now working alongside robots. For a recent survey on the diffusion of robots, AI, and other advanced technologies in the United States, see Nikolas Zolas,

Zachary Kroff, Erik Brynjolfsson, et al., "Advanced Technologies Adoption and Use by U.S. Firms: Evidence from the Annual Business Survey," NBER Working Paper No. 28290 (Cambridge, Mass.: National Bureau of Economic Research, 2020).

[5] Apologies to Arthur C. Clarke.

[6] See, for example, Daniel Zhang, Saurabh Mishra, Erik Brynjolfsson, et al., "The AI Index 2021 Annual Report," arXiv (2021), esp. chap. 2, https://arxiv.org/abs/2103.06312. In regard to image recognition, see, for instance, the success of image recognition systems in Olga Russakovsky, Jia Deng, Hao Su, et al., "Imagenet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision* 115 (3) (2015): 211–252. A broad array of business application is discussed in Erik Brynjolfsson and Andrew McAfee, "The Business of Artificial Intelligence," *Harvard Business Review* (2017): 3–11.

[7] See, for example, Hubert Dreyfus, *What Computers Can't Do* (Cambridge, Mass.: MIT Press, 1972); Nils J. Nilsson, "Human-Level Artificial Intelligence? Be Serious!" *AI Magazine* 26 (4) (2005): 68; and Gary Marcus, Francesca Rossi, and Manuela Veloso, "Beyond the Turing Test," *AI Magazine* 37 (1) (2016): 3–4.

[8] Nilsson, "Human-Level Artificial Intelligence?" 68.

[9] John Searle was the first to use the terms strong AI and weak AI, writing that with weak AI, "the principal value of the computer . . . is that it gives us a very powerful tool," while strong AI "really is a mind." Ed Feigenbaum has argued that creating such intelligence is the "manifest destiny" of computer science. John R. Searle, "Minds, Brains, and Programs," *Behavioral and Brain Sciences* 3 (3) (1980): 417–457.

[10] However, this does not necessarily mean living standards would rise without bound. In fact, if working hours fall faster than productivity rises, it is theoretically possible, though empirically unlikely, that output and consumption (other than leisure time) would fall.

[11] See, for example, Robert M. Solow, "A Contribution to the Theory of Economic Growth," *The Quarterly Journal of Economics* 70 (1) (1956): 65–94.

[12] See, for example, Daron Acemoglu, "Directed Technical Change," *Review of Economic Studies* 69 (4) (2002): 781–809.

[13] See, for instance, Erik Brynjolfsson and Andrew McAfee, *Race Against the Machine: How the Digital Revolution Is Accelerating Innovation, Driving Productivity, and Irreversibly Transforming Employment and the Economy* (Lexington, Mass.: Digital Frontier Press, 2011); and Daron Acemoglu and Pascual Restrepo, "The Race Between Machine and Man: Implications of Technology for Growth, Factor Shares, and Employment," *American Economic Review* 108 (6) (2018): 1488–1542.

[14] For instance, the real wage of a building laborer in Great Britain is estimated to have grown from sixteen times the amount needed for subsistence in 1820 to 167 times that level by the year 2000, according to Jan Luiten Van Zanden, Joerg Baten, Marco Mira d'Ercole, et al., eds., *How Was Life? Global Well-Being since 1820* (Paris: OECD Publishing, 2014).

[15] For instance, a majority of aircraft on U.S. Navy aircraft carriers are likely to be unmanned. See Oriana Pawlyk, "Future Navy Carriers Could Have More Drones Than Manned Aircraft, Admiral Says," Military.com, March 30, 2021. Similarly, companies like Kittyhawk have developed pilotless aircraft ("flying cars") for civilian passengers.

[16] Loukas Karabarbounis and Brent Neiman, "The Global Decline of the Labor Share," *The Quarterly Journal of Economics* 129 (1) (2014): 61–103; and David Autor, "Work of the Past, Work of the Future," NBER Working Paper No. 25588 (Cambridge, Mass.: National Bureau of Economic Research, 2019). For a broader survey, see Morgan R. Frank, David Autor, James E. Bessen, et al., "Toward Understanding the Impact of Artificial Intelligence on Labor," *Proceedings of the National Academy of Sciences* 116 (14) (2019): 6531–6539.

[17] Daron Acemoglu and David Autor, "Skills, Tasks and Technologies: Implications for Employment and Earnings," *Handbook of Labor Economics* 4 (2011): 1043–1171.

[18] Seth G. Benzell and Erik Brynjolfsson, "Digital Abundance and Scarce Architects: Implications for Wages, Interest Rates, and Growth," NBER Working Paper No. 25585 (Cambridge, Mass.: National Bureau of Economic Research, 2021).

[19] Prasanna Tambe, Lorin Hitt, Daniel Rock, and Erik Brynjolfsson, "Digital Capital and Superstar Firms," Hutchins Center Working Paper #73 (Washington, D.C.: Hutchins Center at Brookings, 2021), https://www.brookings.edu/research/digital-capital-and-superstar-firms.

[20] There is some evidence that capital is already becoming an increasingly good substitute for labor. See, for instance, the discussion in Michael Knoblach and Fabian Stöckl, "What Determines the Elasticity of Substitution between Capital and Labor? A Literature Review," *Journal of Economic Surveys* 34 (4) (2020): 852.

[21] See, for example, Tyler Cowen, *Average Is Over: Powering America beyond the Age of the Great Stagnation* (New York: Penguin, 2013). Or more provocatively, Yuval Noah Harari, "The Rise of the Useless Class," Ted Talk, February 24, 2017, https://ideas.ted.com/the-rise-of-the-useless-class/.

[22] Anton Korinek and Joseph E. Stiglitz, "Artificial Intelligence and Its Implications for Income Distribution and Unemployment," in *The Economics of Artificial Intelligence*, ed. Ajay Agrawal, Joshua Gans, and Avi Goldfarb (Chicago: University of Chicago Press, 2019), 349–390.

[23] Erik Brynjolfsson and Andrew McAfee, "Artificial Intelligence, for Real," *Harvard Business Review*, August 7, 2017.

[24] Robert D. Putnam, *Our Kids: The American Dream in Crisis* (New York: Simon and Schuster, 2016) describes the negative effects of joblessness, while Anne Case and Angus Deaton, *Deaths of Despair and the Future of Capitalism* (Princeton, N.J.: Princeton University Press, 2021) documents the sharp decline in life expectancy among many of the same people.

[25] Simon Smith Kuznets, *Economic Growth and Structure: Selected Essays* (New York: W. W. Norton & Co., 1965).

[26] Friedrich August Hayek, "The Use of Knowledge in Society," *The American Economic Review* 35 (4) (1945): 519–530.

[27] Erik Brynjolfsson, "Information Assets, Technology and Organization," *Management Science* 40 (12) (1994): 1645–1662, https://doi.org/10.1287/mnsc.40.12.1645.

[28] For instance, in the year 2000, an estimated 85 billion (mostly analog) photos were taken, but by 2020, that had grown nearly twenty-fold to 1.4 trillion (almost all digital) photos.

[29] Andrew Ng, "What Data Scientists Should Know about Deep Learning," speech presented at Extract Data Conference, November 24, 2015, https://www.slideshare.net/ExtractConf/andrew-ng-chief-scientist-at-baidu (accessed September 9, 2021).

[30] Sanford J. Grossman and Oliver D. Hart, "The Costs and Benefits of Ownership: A Theory of Vertical and Lateral Integration," *Journal of Political Economy* 94 (4) (1986): 691–719; and Oliver D. Hart and John Moore, "Property Rights and the Nature of the Firm," *Journal of Political Economy* 98 (6) (1990): 1119–1158.

[31] Erik Brynjolfsson and Andrew Ng, "Big AI Can Centralize Decisionmaking and Power. And That's a Problem," MILA-UNESCO Working Paper (Montreal: MILA-UNESCO, 2021).

[32] "Simon Electronic Brain–Complete History of the Simon Computer," History Computer, January 4, 2021, https://history-computer.com/simon-electronic-brain-complete-history-of-the-simon-computer/.

[33] Hans Moravec, *Mind Children: The Future of Robot and Human Intelligence* (Cambridge, Mass.: Harvard University Press, 1988).

[34] Will Knight, "Alpha Zero's 'Alien' Chess Shows the Power, and the Peculiarity, of AI," *Technology Review*, December 2017.

[35] Richard Waters, "Techmate: How AI Rewrote the Rules of Chess," *Financial Times*, January 12, 2018.

[36] Matt Beane and Erik Brynjolfsson, "Working with Robots in a Post-Pandemic World," *MIT Sloan Management Review* 62 (1) (2020): 1–5.

[37] Timothy Bresnahan and Robert J. Gordon, "Introduction," *The Economics of New Goods* (Chicago: University of Chicago Press, 1996).

[38] David Autor, Anna Salomons, and Bryan Seegmiller, "New Frontiers: The Origins and Content of New Work, 1940–2018," NBER Preprint, July 26, 2021.

[39] David Killock, "AI Outperforms Radiologists in Mammographic Screening," *Nature Reviews Clinical Oncology* 17 (134) (2020), https://doi.org/10.1038/s41571-020-0329-7.

[40] Erik Brynjolfsson, Tom Mitchell, and Daniel Rock, "What Can Machines Learn, and What Does It Mean for Occupations and the Economy?" *AEA Papers and Proceedings* (2018): 43–47.

[41] Erik Brynjolfsson, Daniel Rock, and Prasanna Tambe, "How Will Machine Learning Transform the Labor Market?" *Governance in an Emerging New World* (619) (2019), https://www.hoover.org/research/how-will-machine-learning-transform-labor-market.

[42] Paul Milgrom and John Roberts, "The Economics of Modern Manufacturing: Technology, Strategy, and Organization," *American Economic Review* 80 (3) (1990): 511–528.

[43] See Daron Acemoglu, Andrea Manera, and Pascual Restrepo, "Does the U.S. Tax Code Favor Automation?" *Brookings Papers on Economic Activity* (Spring 2020); and Daron Acemoglu, ed., *Redesigning AI* (Cambridge, Mass.: MIT Press, 2021).

[44] This reverses the classic result suggesting that taxes on capital should be lower than taxes on labor. Christophe Chamley, "Optimal Taxation of Capital Income in General Equilibrium with Infinite Lives," *Econometrica* 54 (3) (1986): 607–622; and Kenneth L. Judd, "Redistributive Taxation in a Simple Perfect Foresight Model," *Journal of Public Economics* 28 (1) (1985): 59–83.

[45] Tambe et al., "Digital Capital and Superstar Firms."

[46] Katherine S. Newman, *Chutes and Ladders: Navigating the Low-Wage Labor Market* (Cambridge, Mass.: Harvard University Press, 2006).

[47] While the distinction between complements and substitutes is clear in economic theory, it can be trickier in practice. Part of the appeal of broad training and/or tax incentives, rather than specific technology mandates or prohibitions, is that they allow technologies, entrepreneurs, and, ultimately, the market to reward approaches that augment labor rather than replace it.

[48] See David H. Autor, David Dorn, and Gordon H. Hanson, "The China Shock: Learning from Labor-Market Adjustment to Large Changes in Trade," *Annual Review of Economics* 8 (2016): 205–240.

AI, Great Power Competition & National Security

Eric Schmidt

Breakthroughs in AI are accelerating global commercial competition and transforming the international security environment. The reach and influence of foreign-based network platforms present risks to American society and require us to confront questions about their origin and purpose. Meanwhile, AI technologies are enhancing several existing national security threats, and will change the way states try to gain leverage against adversaries and exercise coercion and influence in other societies. The open nature of free and democratic societies, combined with their increasing reliance on poorly secured digital networks, makes them especially vulnerable. In the military realm, AI holds the prospect of augmenting cyber, conventional, and nuclear capabilities in ways that make security relationships among rivals more challenging to predict and maintain, and conflicts more difficult to limit. Even as they compete, rivals should explore limits on AI capabilities. The AI ecosystems of the principal global competitors, the United States and China, remain intertwined, and a calibration of the bilateral technology relationship requires both selective decoupling and continued collaboration in areas of mutual interest. These changes require a comprehensive national strategy for the next decade that preserves global leadership advantages for America's economy and security.

The second decade of the twenty-first century featured two major developments that, together, are shaping the third decade we have now entered. The geopolitical landscape is marked by intensifying competition between the United States and its major power rivals, China and Russia. At the same time, the scientific landscape is characterized by significant advances in artificial intelligence, which promise tremendous economic and strategic advantages for those who capitalize on them.

The confluence of these trends has set up an intense commercial competition among the world's leading technology companies, most of which are based either in the United States or in China. AI is transforming almost every sector of national economies and is accelerating globalized competitions among digital platforms and services. As a consequence, the stakes for future prosperity and long-term national competitiveness are immense.

© 2022 by Eric Schmidt
Published under a Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) license
https://doi.org/10.1162/DAED_a_01916

The security environment is also undergoing significant transformations. This is true across a broad spectrum of national and international security problems, extending from lower- to higher-level intensities of conflict. At the low end, AI is exacerbating cyber and disinformation threats and is changing the way states exercise targeted coercion against opponents. In the middle of the spectrum, warfare between conventional armed forces will feature more rapid actions and delegated decision-making that could make conflict harder to control. At the high end, AI-enabled military and intelligence capabilities may disrupt the fundamental premises of nuclear deterrence in ways that undermine strategic stability.

All of this requires a comprehensive national strategy for the next decade that preserves global leadership advantages for both America's economy and security. The United States must protect against hacking, coercion, and other efforts by adversaries to use our society's openness against us. The most dangerous aspects of the U.S.-Chinese and U.S.-Russian military rivalries must be managed to avoid disastrous conflicts. And the innovation economy that has put American technology and ingenuity at the forefront of scientific advances for decades must be bolstered to stay ahead of America's principal competitor, China.

Understanding how AI drives the new global commercial landscape begins with *network platforms*, which I describe in a recently published book, *The Age of AI*, coauthored with Henry Kissinger and computer scientist Daniel Huttenlocher, as digital services that provide value to their users by aggregating them in large numbers, often at a transnational and global scale.[1] Today, the major network platforms increasingly rely on AI for growth. A network platform's value and attractiveness grow as additional users adopt it. The potential social, economic, political, and geopolitical influence of each major network platform is substantially augmented by the degree of these positive network effects.

Two features of global network platforms are especially significant to geopolitics. First is their tendency toward consolidation. As more users are drawn to certain platforms, their network advantages reduce competition, leaving us with fewer providers of a given service, each with a large base of users. In other words, the dynamics of positive network effects tend to support only a handful of major players who are operating at the forefront for their product or service. The small number of leading platforms thereby gain and exercise significant influence on a global scale.

The second feature is that many nations are, and are likely to remain, reliant on network platforms that are both designed and hosted in other countries. As a result, they are dependent, at least in part, on other countries' regulators for continued access, key inputs, and international updates. In the United States and elsewhere, this has created concerns about the implications of conducting broad aspects of national economic and social life on network platforms that were built

in other, potentially rival, countries. These platforms may foster a close level of connection and influence, particularly with the use of AI to learn from and steer a country's citizens.

Taken together, these two features point to a growing geopolitical and national security concern for the United States. The globally dominant network platforms of the future could be based in rival countries and could exert significant influence over American society and even critical infrastructure. If a network platform is useful and successful, it comes to support broader commercial and industrial functions and, in this capacity, may become nationally indispensable. At least theoretically, the threatened withdrawal of such a network platform serves as a potential instrument of leverage. This hypothetical ability to "weaponize" network platforms by withholding service in a crisis is an increasingly significant factor in national security planning.

The reach and influence of global network platforms require us to ask essential questions about their origin and purpose: By whose design, and with what regulatory parameters, is the AI operating? What impact might these processes have on social norms and institutions? Who has access to the information generated through the platform?

Looking across the world, a multidisciplinary contest for economic advantage, digital security, technological primacy, and ethical and social objectives is unfolding.

The United States has begun to view network platforms as an aspect of international strategy, restricting the domestic activities of some foreign platforms and restricting the export of some software and technology that could strengthen foreign competitors. At the same time, critics in and out of government have identified major domestic network platforms as targets for antitrust actions. This simultaneous drive for strategic preeminence and domestic multiplicity may push U.S. development in opposing directions.

Meanwhile, China has similarly supported the development of formidable network platforms that are global in scale and poised to expand their reach. Beijing has also taken steps to shape international technology standards and bar the export of sensitive, domestically developed technologies. Chinese network platforms predominate in China and nearby regions, and some are leading global markets.

The landscape will also be shaped by actors beyond the United States and China. Europe, unlike China and the United States, has yet to create homegrown global network platforms or cultivate the technology industry that has supported the development of network platforms elsewhere. To be sure, Europe has a significant place in the global AI landscape with some leading companies and universities, sizable markets, and a formidable regulatory apparatus. Yet Europe continues to face disadvantages for the initial scaling of new network platforms due to the

many languages and separate national regulatory systems in Europe's combined market.

The European Union has focused recent regulatory attention on obliging changes in American and (to some extent) Chinese network platforms' conduct as a condition of their operation in the European market. Europe faces the choice of whether to act as an ally to one side or another in each technological sphere or to act as a balancer between sides. Here, the preferences of the traditional, core EU states and the newer Central and Eastern European entrants may differ, reflecting varying geopolitical and economic experiences. Thus far, historic global powers like France and Germany have prized independence and freedom to maneuver in their technology policy, whereas more peripheral European states with recent and direct experience of foreign threats (such as the post-Soviet states) have shown greater readiness to identify with a U.S.-led technology sphere.

While still an emerging force in this arena, India has substantial intellectual capital, a relatively innovation-friendly business and academic environment, and a vast reserve of technology and engineering talent that could support the creation of leading network platforms. India's population and economy are of a scale that could potentially sustain independent network platforms, without recourse to other markets. Likewise, Indian-designed network platforms have the potential to become popular in other markets as well. As India assesses its regional relationships and relative reliance on imported technology, it may elect either to chart a more independent path or to assume a principal role within an international bloc of technologically compatible nations.

The Global AI Index, the most comprehensive effort to date to rank countries in terms of AI advancement, offers several insights into how the global competitors stack up.[2] The creators of the index assessed countries based on 143 indicators across areas such as the talent of AI researchers and practitioners, infrastructure, R&D, government strategy, and commercial activity. This is, of course, a snapshot in time. What emerges, though, is the centrality of AI talent indicators to assess both current strength and future trends. Consider, for example:

- The United States leads China by the widest margin in the talent category (scoring five times higher). It also holds significant leads in research and in commercial AI. These factors seem naturally related: the best talent is producing the best research and driving the best commercial products. All of this points to the importance of keeping America's global edge in attracting and retaining top AI talent.
- Several Western allies also score higher than China in AI talent, including the United Kingdom, Canada, Germany, the Netherlands, and France. This raises questions about the extent to which European states will be able to capitalize

on this excellent talent base after Brexit and in the midst of the EU's evolving regulatory approach to AI.

- Although India's overall score is much lower than China's, its talent score is substantially higher. India ranks second in the world in AI talent, behind only the United States. This suggests tremendous potential for India to emerge as a global AI heavyweight over time, if India can improve its position in other areas such as national infrastructure, government strategy, and commercial application.

In overall scores, the United States and China are in a league of their own at number one and two, respectively. But the next tier (about ten countries whose overall scores fall in the same ballpark) is made up entirely of U.S. allies and partners. This points to the critical need to develop and strengthen AI partnerships with those nations.

Depending on how the commercial competition unfolds – even with, or perhaps as an effect of, such global partnerships – an industry founded on the premise of global community and communication may, in time, be enlisted in a process of regionalization. Such a process could unite blocs of users in separate realities, influenced by distinctive AI that has evolved in different directions and is shaped by spheres of regional technology standards. While these trends play out, some of these AI-driven platforms will be at the center of novel national security challenges.

Artificial intelligence technologies are enhancing several existing national security threats and will change the way states try to gain leverage against adversaries and exercise coercion and influence in other societies. The open nature of free and democratic societies, combined with their increasing reliance on poorly secured digital networks, makes them especially vulnerable to these threats.

In its 2021 final report, the National Security Commission on Artificial Intelligence, an independent government panel that I chaired, found that the machine learning algorithms that transformed how business was done in the early years of this century are now transforming intelligence and statecraft.[3] Technology and advertising companies learned the value of AI for harvesting and analyzing consumer data. Similar capabilities wielded by governments can now be used for espionage, targeted coercion, tailored disinformation, sophisticated cyber intrusions, and potentially biological warfare.

AI opens new horizons of capabilities in the information space, both in monitoring and in disinformation and disruption. In theory, at least, AI could be used to determine the most effective ways of delivering synthetic content to people, tailoring it to their biases and expectations. Both "offense" and "defense" – both

the spread of disinformation and efforts to combat it – will become increasingly automated and entrusted to AI.

These capabilities could be used across the spectrum of conflict: as tools of pressure during peacetime, as a prelude to military actions, or in concert with a military campaign.

One implication of these changes is that data security has become a more central problem of national security. AI makes it harder to protect personal information – finances, patterns of daily life, relationships, and health among other things – that adversaries could use to develop individually tailored models for influence. This is the major counterintelligence challenge for the AI era.

Another, related security concern is that the cyber domain is becoming increasingly complex and automated. Once AI-enabled malware is lodged onto a computer system, it will be able to mutate into multiple forms to avoid detection and countermeasures. Such mutating polymorphic malware already accounts for the vast majority of malicious executable files circulating in cyberspace.

The U.S. government's tools to manage these threats are clearly inadequate. Substantial changes are required in the way we think about data security and in our policies and laws to strengthen it. We need to identify categories and combinations of our most sensitive personal and commercial data, and develop a broad approach with clear policies, criteria, or authorities to confront this multifaceted problem. Likewise, major reforms are needed in cybersecurity, including widespread integration of AI-enabled cyber defenses to match and neutralize offensive AI-cyber techniques.

The AI era risks complicating the riddles of modern strategy beyond human intention, or perhaps even human comprehension. AI holds the prospect of augmenting cyber, conventional, and nuclear capabilities in ways that make security relationships among rivals more challenging to predict and maintain, and conflicts more difficult to limit.

AI's capacity for autonomy and logic generates a layer of incalculability. Most traditional military strategies and tactics are based on the assumption of a human adversary whose conduct and decision-making calculus fit within a recognizable framework or have been defined by experience and conventional wisdom. Yet an AI system piloting an aircraft or scanning for targets follows its own logic, which may be inscrutable to an adversary and unsusceptible to traditional signals or feints and which will, in most cases, proceed faster than the speed of human thought.

Moreover, because AIs are dynamic and emergent, even those powers creating or wielding an AI-designed or AI-operated weapon may not know exactly how powerful it is, or what it will do in a given situation. When actors deploy AI weapons against one another, neither side may have a precise understanding of what their interaction will generate or what may be its collateral effects.

The integration of AI into military and intelligence systems heightens the risk of instability and conflict between the United States and its rivals across a spectrum of scenarios, from activities beneath the threshold of war, to conventional warfare between armed forces, to nuclear escalation.

At the lower end, for example, it is not hard to imagine how AI-enabled capabilities could provide China with more effective tools to patrol the South China Sea and consolidate its strategic position there. Nor is it hard to imagine Russian cyber and disinformation activities in Ukraine or elsewhere in Europe becoming more effective, persistent, and influential with AI.[4]

Once they are released into the wild, AI-enabled cyber weapons may be able to adapt and learn and may go well beyond their intended targets. The very capabilities of the weapon might change as the AI reacts to its surroundings. The multibillion-dollar global damage caused by Russia's 2017 NotPetya attack concretely demonstrates the power of even basic automated malware, the risk tolerance of capable state actors, and the consequences of such capabilities proliferating.

AI-enabled cyber weapons may allow adversaries to launch digital assaults with exceptional speed, dramatically accelerating the human capacity to exploit digital vulnerabilities. As such, a state may effectively have no time to evaluate the signs of an incoming attack. Instead, they may need to respond immediately or risk disablement. If they have the means, they may elect to "respond" nearly simultaneously, before the event can occur, constructing an AI-enabled system to scan for attacks and empowering it to counterattack. This could lead to new forms of automated preemption or anticipatory self-defense and strain the legal and policy frameworks that guide government decision-making.

In conventional warfare, greater reliance on automated capabilities, combined with the intense decision-making time pressures that attend operations conducted at machine speeds, could lead to rapid and even unintended escalation. This is all the more worrisome if militaries rush to field new systems that are unreliable in practice and poorly understood by operators. Unintended escalation could occur for many reasons – including when systems fail to perform as intended because of interactions between opposing systems on the battlefield, or as the result of machines or humans misperceiving signals or actions. As AI-enabled systems increase the pace of warfare across the board, the time and space available for de-escalatory measures will shrink.

There are also reasons to believe AI will erode nuclear stability, although some of these concerns are largely theoretical for now. For example, if AI-enabled intelligence and targeting systems are better able to locate nuclear forces that are currently hard to see and strike (because they are under the sea or moving around on land), this would put at greater risk a state's second-strike capability and thereby undermine mutual vulnerability, which is considered to be a source of stable nu-

clear deterrence. Other concerns relate to potential integration of AI into nuclear command and control.

In the military sphere, realism should compel rivals, even as they compete, to explore limits on the development and use of certain destructive, destabilizing, or unpredictable AI capabilities. This could include a sober effort at some form of AI arms control or, if that is too ambitious, the development of confidence-building measures between rival states to reduce risks to international stability.[5]

If weapons can change in ways that prove different in scope or kind from what their creators anticipated or threatened, calculations of deterrence or escalation have the potential to turn illusory. Moreover, from a technical standpoint, the lines between engaging AI in reconnaissance, targeting, and lethal autonomous action may be relatively easily crossed, making a search for mutual restraint and verification systems difficult but imperative.

To be meaningful, restraints must be reciprocal. But the management of mutual restraints on military AI systems will be even more difficult than it has been for nuclear weapons, which has been the endeavor of more than a half century of diplomacy among rivals and remains incomplete and fragmentary. The challenge of assessing the nuclear balance is relatively straightforward. Warheads themselves can be counted and their yields known. Conversely, the capabilities of AI are not fixed, they are dynamic. Unlike nuclear weapons, AI systems are hard to track: once trained, they can be copied easily and run on relatively small machines. And detecting their presence or verifying their absence is difficult or impossible with present technology. This is an important area for further technical research and policy development.

To begin approaching these questions through diplomacy, initial U.S. dialogue with China or Russia should focus on making sure that both sides know, at least in general terms, what the other is doing. Such a discussion of AI weapons among major powers must be endeavored, if only to develop a common vocabulary of strategic concepts and some sense of each other's red lines.

Because the incorporation of AI systems in nuclear strategy is still nascent, now is the window of time for nuclear states to discuss protocols and understandings that could minimize the disruption to nuclear stability. One helpful measure would be to clearly and publicly affirm existing U.S. policy that only humans can authorize the employment of nuclear weapons – and then seek similar commitments from other states.

At the same time, the United States and other major powers should make efforts to limit the proliferation of AI-enabled weapons. Once introduced, these capabilities could spread quickly. Although creating a sophisticated AI requires substantial computing power, proliferating the AI or running inference generally

does not. AI will be ubiquitously acquired, mastered, and employed; the imposition of restraints on weaponizing AI, or even achieving a collective definition of restraint, will be exceedingly difficult.

Security risks to the United States will become more acute if China's researchers, companies, and military and intelligence agencies overtake their American counterparts in AI proficiency and breakthroughs. At the same time, an open international research environment encourages mutually beneficial scientific advances in both countries. Adjusting the degrees to which U.S.-China technology relations should be open or closed will remain an evolving challenge.

Only the United States and China have the resources, commercial might, talent pool, and innovation ecosystem to lead the world in AI. In some areas of AI research and application, China is a peer, and in certain applications, China is already more technically advanced. Within the next decade, China could surpass the United States as the world's preeminent AI power.

If China's firms win the competition for global network platforms, it will not only disadvantage U.S. companies, it will also create the digital foundation for a geopolitical challenge to the United States and its allies. Platform domination abroad allows China to harvest the data of its users and permits China to extend aspects of its domestic system of control. Wherever China controls the digital infrastructure, it will gain greater leverage to conform the world to its goals.

Meanwhile, the research ecosystems in China and the United States are deeply connected through shared research projects, talent circulation, and commercial linkages that include supply chains, markets, and joint research ventures. It would be counterproductive to sever the technology ties to China that foster basic research and benefit U.S. companies. But the United States must safeguard the integrity of open research, prevent the theft of American intellectual property, and employ targeted tools like export controls and investment screening to protect technology industries that are critical to national security.

An appropriate calibration of the U.S.-China technology relationship would include: 1) some purposeful decoupling of specific linkages that introduce unacceptable vulnerabilities, such as in areas with clear security and military applications; 2) continuing cooperative research that brings significant joint benefit; 3) continuing commercial interchange between technology sectors; 4) greater collaboration in shared scientific challenge areas; and 5) increased federal government investment in research and development, which will help position the United States to win network platform competitions.[6]

Decoupling, through this lens, is not just about disconnecting from China. It is about revitalizing America's own productivity in critical areas. At the same time, the United States must also build up the capacity of its allies and partners. Done

right, purposeful decoupling could spur a commercial renaissance in particular classes of technologies across Western nations.

Breakthrough progress by China on several fronts has intensified the U.S.-China technology competition. The United States must continue to invest in American innovation to keep from falling behind. There has been a continuity of purpose across administrations to mount a major national effort in AI. Keeping the momentum requires the federal government to take a more assertive role than Americans have been accustomed to in recent decades.

Most technology advances in the United States will be driven by the private sector and universities. Although publicly funded research has been important for innovation, the private sector has proved to be America's great strength. Companies move faster and more globally than any government could. But large technology firms cannot be expected to compete with the resources of China or make the large, nationwide investments the United States needs to stay ahead in the competition. A hybrid approach that more tightly aligns government and private sector efforts is needed to win.

One example of such an approach is the National AI Research Resource (NAIRR), a recommendation of the AI Commission. Requested by Congress through the National AI Initiative Act of 2020, this initiative aims to democratize access to compute environments, data, and testing facilities, providing researchers beyond the leading industry players and elite universities with the ability to pursue cutting-edge AI work. The initiative promises to spur nationwide technology advances with benefits for overall national competitiveness.

Another area for constructive government action is in microelectronics. After decades of leading the microelectronics industry, the United States is now almost entirely reliant on foreign sources for production of the cutting-edge semiconductors that power the AI algorithms critical to everything from our defense systems to our smartphones. The dependency on semiconductor imports, particularly from Taiwan, creates a strategic vulnerability from adverse foreign government action, natural disaster, or other events that could disrupt supply chains for electronics. At the same time, China has made an enormous financial commitment to forging a world-leading semiconductor industry by 2030, with the goal of minimizing or eliminating China's own dependency on imported microelectronics. The United States must be committed to a strategy to stay at least two generations ahead of China in state-of-the-art microelectronics. Doing so requires continued funding and incentives to maintain multiple sources of cutting-edge microelectronics fabrication in the United States.

In the coming years, economic and security competitions will proceed in parallel, with China aiming to achieve global preeminence in AI by 2030 and security agencies among all competitors adopting AI for a wide range of applications.

Trends in global network platforms will not just define the landscape of commercial AI, but will also shape the security environment in novel ways. International stability will hinge in large measure on whether rival states can manage the uncertainties of AI in the cyber, conventional, and nuclear realms. And the United States will need to carefully navigate its interdependencies with China while also continuing domestic reforms to bolster innovation. How the United States manages these interrelated challenges will go a long way toward determining its competitive position by the end of the decade.

ABOUT THE AUTHOR

Eric Schmidt, a Fellow of the American Academy since 2007, is the former Chief Executive Officer of Google and former Executive Chairman and Technical Advisor of Alphabet, Inc. He is also a Founder of the Schmidt Foundation, the Schmidt Ocean Institute, and Schmidt Futures. He is the Chair of the Special Competitive Studies Project and was the Chairman of the National Security Commission on Artificial Intelligence from 2019 to 2021. He is the author of *The Age of AI: And Our Human Future* (with Henry Kissinger and Daniel Huttenlocher, 2021), *Trillion Dollar Coach: The Leadership Playbook of Silicon Valley's Bill Campbell* (with Jonathan Rosenberg and Alan Eagle, 2019), and *How Google Works* (with Jonathan Rosenberg, 2014). In 2020, he launched the podcast Reimagine.

ENDNOTES

[1] This discussion on network platforms and global commercial competition in AI draws in part from Henry A. Kissinger, Eric Schmidt, and Daniel Huttenlocher, *The Age of AI: And Our Human Future* (Boston: Little, Brown and Company, 2021).

[2] The Global AI Index, Tortoise Media, https://www.tortoisemedia.com/intelligence/global-ai/ (accessed January, 2022).

[3] National Security Commission on Artificial Intelligence, *Final Report* (Washington, D.C.: National Security Commission on Artificial Intelligence, 2021), https://www.nscai.gov/2021-final-report/.

[4] This essay was written before Russia's invasion of Ukraine in February 2022.

[5] This discussion of measures to mitigate international stability risks draws in part from Kissinger et al., *The Age of AI*.

[6] For an elaboration of the argument for calibrating U.S.-China technology relations, and further explanation of these five features, see Eric Schmidt, "Building a New Technological Relationship and Rivalry," in *COVID-19 and Global Order*, ed. Hal Brands and Francis Gavin (Baltimore: Johns Hopkins University Press, 2020).

The Moral Dimension of AI-Assisted Decision-Making: Some Practical Perspectives from the Front Lines

Ash Carter

This essay takes an engineering approach to ensuring that the deployment of artificial intelligence does not confound ethical principles, even in sensitive applications like national security. There are design techniques in all three parts of the AI architecture – algorithms, data sets, and applications – that can be used to incorporate important moral considerations. The newness and complexity of AI cannot therefore serve as an excuse for immoral outcomes of deployment by companies or governments.

One of the most frequent questions I was asked as U.S. Secretary of Defense (2015 – 2017) was whether there will be autonomous lethal weapons. My answer was no, the U.S. Department of Defense (DOD) would not deploy or use truly autonomous systems in the application of lethal force. Being technologically inclined, I established the Pentagon's official policy in a memorandum back in 2012 when I was Deputy Secretary. When conceiving of this directive, I had imagined myself standing in front of the news cameras the morning after innocent bystanders had been killed in an airstrike aimed at terrorists or opposing combatants. And suppose I answered in response to obvious and justified interrogation over responsibility: "It's tragic, but it's not our fault: the machine did it." This reply would be rightly regarded as unacceptable and immoral.

What, then, can ethically "justify" the risk of a terrible error made in the application of artificial intelligence?[1] In one sense, nothing, of course. Yet as a practical matter, AI is going to be used, and in an ever-widening set of applications. So what can bound moral error? Algorithm design? Data set selection and editing? Restricting or even banning use in sensitive applications? Diligent, genuine, and documented efforts to avoid tragedies? To some extent, all of these.[2] The fact that there are practical technical approaches to responsible use of AI is paramount to national defense. AI is an important ingredient of the necessary transformation of the U.S. military's armamentarium to the greater use of new technologies, almost all of them AI-enabled in some way.

© 2022 by Ash Carter
Published under a Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) license
https://doi.org/10.1162/DAED_a_01917

This essay takes a technical rather than a legal approach to AI ethics. It explores some practical methods to minimize and explain ethical errors. It provides some reasons to believe that the good to be obtained by deployment of AI can far outweigh the ethical risks.

The 2012 DOD Directive 3000.09 reads "autonomous and semi-autonomous weapons systems shall be designed to allow commanders and operators to exercise appropriate levels of human judgment over the use of force."[3] This guidance has been reissued and refined several times since.

Note that the directive does *not* use the language "man in the loop." To use such a formulation would be technically ignorant and utterly infeasible. The whole point of the machine is to operate faster, more accurately, and frequently entirely without communication with humans (that is, "autonomously"). Thus, the image of a person inserted in the circuitry like a living chip is ridiculous. In certain ways, the whole idea of autonomy in warfare is not at all new. Take a guided anti-air missile, for example: most of these find their way to their target – during their entire trajectory, or at least in their lethal "endgame" – using inputs from a homing seeker in the nose (a camera, say, or radar) whose output is calculated on board the missile with computers and software designed and tested years previously and updated during flight. In these respects, the question about autonomous weapons or "lethal AI" has thus been around for quite a while. Still, with AI and many of its applications developing at lightning speed, we must give some good answers to its distinctive questions. The language of the directive was crafted to suggest that the DOD would insist upon other more practical forms of "human judgment" built into its AI-enabled weapons systems.

It is not particularly surprising that the tradecraft for ethical use of AI has lagged development of the technology itself. For one thing, after pioneering AI as early as the 1950s, the DOD ended up lagging in its application to military problems (in a manner all too common). It is relatively recent that the Pentagon has begun catching up. And while universities do critical fundamental research, including interdisciplinary work bringing tech together with law, ethics, policy, and other fields of thought, they are remote from direct application at scale. Instead, applied work in AI has been led by the consumer Internet and advertising industries. These industries can afford to be tolerant of Type I errors (pushing content to users unlikely to buy the sponsor's products, for example) or Type II errors (not pushing content to likely customers) in accomplishing their objectives. The analogous kinds of errors would be much graver in applications like national security, health care, self-driving vehicles, or law enforcement. The ethical errors of privacy violations and manipulation that have clearly been made by consumer Internet and advertising companies are not errors of AI, but of a basic lack of moral self-scrutiny. In fact, a significant part of the American creative culture, at least in

digital technology, has believed that its dynamism springs from its independence from government and virtually any meaningful regulation (strikingly illustrated by Section 230 of the Communications Decency Act of 1996, which provided broad immunity to online platforms from civil liability for content on their platforms).

But in spite of tech's preference for a freewheeling environment and government's deserved reputation for stodginess, public policy is irreplaceable for technological progress. Government is the artery through which not only vital basic research funding flows, but also the rules, norms, and regulations that fortify acceptance and trust by the population of technological progress as something that is a net positive for humanity. Historically, the "disruptive" technology-enabled industrial revolution that resulted in the gigantic farm-to-factory migration was only successful in America because the government supplied the complementary ingredients to those supplied by innovators and profit-seeking industry. Government-supplied ingredients like universal public education prepared a farmland workforce for industrial jobs. Progressive-era labor reforms offering legal and other safeguards for workers made it possible for most Americans to support a free market system of large corporations and to view technology as a net positive. In other countries, notably Russia, the farm-to-factory revolution failed. During roughly the same period, U.S. government standards for the safety of foods and drugs were promulgated – and accepted – at a nationwide scale. Without such regulation, citizens would not be able to trust the industrialized production of foods, and there might not have been continent-wide markets like those that developed quickly in America.

In a similar way, it seems likely that AI and ethics will mix best when tech and government mix well. The purely technical challenges of ethical AI are hard enough. We do not need a failure of government and industry collaboration to be an obstacle to the ethical fielding of AI. As a technologist as well as a government leader, I believe strongly in both the wonders of AI *and* the importance of morality in engineering. I am also optimistic this can be solved, like every other hard problem, with diligent technology-informed effort. This will be essential for national defense.

While some advancements in AI are breathtakingly new, their novelty should not be exaggerated. Right and wrong are certainly not new. The question has been around a long time: what is a "good reason" for the rest of us to soften the penalty for, or excuse entirely, the people who designed and sold the technology that made a tragic error? Various justifications have long been defined in courts, product warranties, corporate disclosure statements, and press conferences. There is even a rule-of-reason that recognizes that no technology is perfect, so "good" only means "good enough." Not just the morality but also the political practicality of deploying AI hinge on some sort of accountability and responsibility engineered into it that is "good enough" for the purpose. Ethical design principles can

be identified in all three components of an AI deployment: algorithms, data sets, and applications.

Many kinds of AI algorithms exist in practice, and even more are being developed or hypothesized. They all make enormous numbers of tiny calculations that combine to make overall inferences that cannot be made quickly by humans, have not been recognized by humans, or even perhaps would never be recognized by humans. These computational methods make literal transparency, normally the starting point for ethical accountability, completely impractical. It is usually impossible to "deconvolve" the series of steps leading to the inferences made by AI.

Moreover, like software of nearly any kind, AI algorithms are the product of many hands and many engineers working in many venues over many years. While blame for an unethical outcome can be attributed to the final vendor or end user, even this is unreasonable unless negligence can be shown, which takes us back to the same fundamental dilemmas.

One approach is to make ethics an internal algorithm design criterion from the start. Doing so successfully may require substantial new conceptual invention in its own right, but this can be as exciting for the coding engineers as maximizing any other design feature, especially if value is attributed to it. The federal government, including the DOD, should fund basic research into ethics-by-algorithms, recognizing that companies will underinvest until some terrible wrong occurs. My experience in technology management suggests that the initial specialist refrain "it can't be done" is usually overcome by making the desired innovation a requirement-to-buy or a weighted factor in competitive source selection.

An additional approach is to focus on the *process* of algorithm design rather than the algorithm itself. The history of processes designed to prevent the misuse of nuclear weapons offers a valuable example. Bombs themselves are outfitted with elaborate coded locks to prevent abuse, which could have the gravest consequences. But any repair, movement, or contact – that is, any process in which bombs are handled, moved, repaired, or altered – requires two people rated in the same specialty (the "two-man rule"). Even I, as Secretary of Defense, was not authorized to be alone with a nuclear weapon. These many simultaneous approaches to security policy, some involving design and some involving process, recognized the ineffable variety of possible failure modes and the absolute necessity to prevent every one of them, all in an essentially unending custodianship of tens of thousands of bombs (the half-life of Plutonium-239 is twenty-two thousand years and Uranium-235 is 703 million years). The complexity of this challenge was deepened with the collapse of the Soviet Union, which fragmented the systems that had served to control one of the world's two biggest collections of such weapons. In the 1990s, I ran the Pentagon program created to assist the post-Soviet militar-

ies to protect and reduce the arsenals they inherited, and I was in awe of the way such a combination of design and process methods safeguarded weapons in a totally unforeseen social disintegration.

Programs of established design principles backed up by dual or multiple checkers with equal training and qualifications and redundant safeguards are widely used in complex systems. Establishing such a design process control can not only reduce the likelihood of errors with advanced AI applications but mitigate, at least partially, the liability assigned to innovators if they do occur. It was precisely such a process protocol that was apparently compromised in the famous case of the Boeing 737-Max. Its in-flight controls were reportedly the cause of two back-to-back airline crashes. The Federal Aviation Administration supposedly failed to provide thorough expert checking of the significant changes to in-flight characteristics that occurred when the older 737 was changed to the Max configuration. Among the fatal mistakes was the sacrifice of an established design criterion in the software itself requiring dual redundant sensor inputs to the fly-by-wire flight controls. Due to the COVID epidemic, most people are by now familiar with the Food and Drug Administration's "safety and efficacy" testing that must precede release of a new vaccine. So the notion of requiring a process of qualified review for sensitive products is hardly new and should be the industry standard for AI.

A dilemma arises from proprietary secrecy. A vendor will not want to disclose the inner workings of its algorithms and data sets; these are sensitive for competitive reasons. Given proprietary concerns, it is advantageous to establish industry-wide standards and a level of government involvement in the certification that these standards are being met. Government routinely handles proprietary secrets of competing companies when it serves as a regulator or customer of advanced technology. Government security classification sometimes can be argued to slow the pace of innovation by preventing the free flow of ideas. But in the case of most AI, the preponderance of innovation is centered in companies, and intercompany secrecy is by far and away the bigger barrier to sharing information, the more so as the research frontier has moved out of universities that publish results openly and into industry.

It is worth noting that AI itself can be a powerful tool in certification testing of AI systems whose workings are impossible for humans to fully grasp. The "checking AI" can perform an exhaustive search for oddities in large numbers of input-output runs and thereby identify design defects without unpacking the full mass of layered calculations. In the same way, AI can conduct cyber defense by probing randomly around the victim's attack surface for unidentified holes, simulating the "rat in a maze" attack (to distinguish it from the attacker who exploits exquisite defects discovered in the victim's defenses – the "jewel thief"). This is just a new case of an old pattern in technology and warfare, in which the same invention that creates new dilemmas can also help protect from those very dilem-

mas. AI-assisted checking of algorithms can also speed up the process of ethical audit so it does not delay deployment.

The next thing to tackle in ethical AI is the data set the algorithm is trained on (if it is machine learning) or otherwise crunches to make recommendations to the human user. Data sets come from a wide variety of huge caches: enterprise business systems, social media, search engines, public data sets, the entire historical corpus of the written word, and Internet of Things (IoT) and sensor data of all kinds. The trickiest sets are "unstructured data": impressively large jumbles of data collected in an incidental manner.

Generally, it really is true: "garbage in, garbage out." Some open-mindedness is needed, however, in the case of AI. Important hints or suggested solutions might come from running on bad data, but they should not be used for making determinations in sensitive applications.

An "ethical audit" of an AI database begins with its provenance. It seems well established that true anonymity cannot be promised: AI is so thoroughly penetrating that individual identity can almost always be unwound. It turns out that the risk of identification goes up in surprising ways when two databases, assembled "anonymously," are combined. There are technical approaches to enhancing privacy and true anonymity in databases used by AI that seem durable. One example is provided by the various forms of "differential privacy" in which fake data are mixed with true data in a quantified way, preserving some privacy but not entirely spoiling the data's use. The Census Bureau uses differential privacy in its data.

It is also clear that "informed consent" is not a good ethical proxy in data collection and exploitation without expert guardrails. Few of us can really understand on our own the full consequences of our consent. A company selling or deploying AI that abuses personal data should not be able to evade responsibility by citing the supposedly informed consent of the victims.

Some data sets are morally questionable from the start, for example those collected in communist China for purposes of dictatorship and control. It is often said that China will outperform the United States in AI because its population of 1.3 billion, or three times the United States', provides a database size advantage. But I am unaware of any design or implementation of AI that is qualitatively better because of a factor of three in data set size. The real difference is the intrusive methods of Chinese data collection and application. China is indeed likely to excel in the AI of totalitarianism, but this is hardly enviable from an ethical perspective.

Assuming that the data sets used in AI are collected ethically to begin with, three features need to be carefully audited for inaccuracies and biases that could lead to morally fateful events when they are deployed. The audit should encompass the training set (in the case of machine learning), the application set, and

potential issues in matching the two. As in the case of algorithms, an ethical case can be built on the characteristics of the data themselves and the process by which they are audited.

To my knowledge, there is no substitute for a qualitative examination with a skeptical eye. Is the entire space of possible data points defined and is there a reasonable presence (or understood absence) of points in some corners (such as an edge subset representing a minority)? Is the set examined against a checklist of possible flaws: biased, outdated, or otherwise unrepresentative? How were the data originally tagged? Way back in the provenance of most data sets is a human tagger who originally assigned a location to each point in a dataspace ("is this a dog or cat?"). And again, as in the case of algorithms, AI itself can help work through a proposed data set against a checklist of possible foibles before deployment. Finally, and again as with algorithms, the process of database audit itself can be given ethical standards: documentation, multiple qualified checkers, simulations, and sampling.

The application is the last ingredient in the consideration of ethics in AI. Strong ethical efforts with algorithms and data sets of the kind discussed above are not really needed in some applications. Entertainment and advertising, as already noted, can be fairly error tolerant. It is up to the user or customer. But there are applications that require much more ethical scrutiny: national security, of course (and especially the use of force); law enforcement; health care; autonomous vehicles of all kinds; fairness in credit, housing, and employment; and at least some parts of elections and political life. An "in-between" category might be commerce and some parts of finance, where the risk of error is mostly economic rather than moral and can be priced in. And even seemingly innocent applications in the consumer Internet can turn in dark directions when their true mission is deceit, manipulation, privacy violation, or their enablement.

The reason the techniques for scrutiny of AI algorithms and data sets described above are important is that the complexity and relative newness of AI can conceal ethical problems from even ethical users of technology. In this respect, AI is no different from other new technologies: they always create new capabilities that must be situated in a framework of right and wrong that itself changes slowly, or arguably not at all. Nuclear weapons, for example, certainly created new capabilities of mass destruction, but the moral principles of just war, proportionality, and discrimination still applied to them.

I believe that this discussion of AI algorithms and big data sets demonstrates that AI is not impenetrable. It is possible to locate right and wrong even in AI's amazing complexity. It is *not* possible to claim that the technology itself makes moral use indefinable. It even follows that occasional tragic outcomes are defensible if these techniques have been used with care. What is indefensible is applica-

tion of AI to inherently immoral purposes, or deployment without the technical efforts described above.

I cannot rule out that someday AI might create truly new ethical puzzles for humanity. Such dilemmas would have to evade both deeply informed expert scrutiny of the technology and extant moral principles. While the popular press sometimes alleges that AI has created qualitatively new ethical quandaries, no such cases have been found to date.

There is, therefore, a variety of engineering approaches to building ethics into AI. The technical judgment that AI and big data, despite their seemingly ineffable complexity, do not defy moral examination is good news for U.S. national defense. As Secretary of Defense, I made no apology for the fact that America takes its values to the battlefield. But I also made a large number of changes in the DOD's structure and practices to connect the Pentagon more closely to the tech sector. When I took my first job in the Pentagon in 1980, most new technology, including AI, originated in America, and most of that under government (largely under DOD) sponsorship. Today, the tech base is commercial and global. For this new era, the Pentagon therefore needs to build new and different kinds of bridges to the tech sector. Accordingly, as Secretary of Defense, I founded the Defense Digital Service to bring young techies directly into the Pentagon, placed Defense Innovation Unit outposts in the nation's tech hubs, and convened a Defense Innovation Board chaired by Google's Eric Schmidt.

But in the same role I also authorized raids, hostage rescues, counterterrorist operations, ongoing combat operations in Afghanistan, the major campaign to destroy the Islamic State, and a host of war plans devised for China, Russia, Iran, and North Korea, all of them requiring grave moral judgments and all of them using the newest technology the Pentagon had. It is important that leaders be able to situate important moral principles in dramatically new technological settings, rather than being bamboozled into thinking they do not apply.

The list of exploding tech fields is long. It encompasses all forms of intelligence collection and electronic warfare, cyber warfare, robotic vehicles, ubiquitous presence via space, IoT, global WiFi and LiFi, bioengineering and biodefense, all sorts of new engineered materials, undersea warfare, microsatellites, human performance enhancement, various quantum applications, directed-energy weapons, and hypersonic vehicles. To make room for these innovations in the defense budget, familiar military capital stock like manned armored vehicles, many surface ships and large-mass satellites, manned aircraft, and even certain infantry subspecialties will gradually be phased out. The only field of warfare in which changes are not anticipated is nuclear weapons. Without exception, each of the new technologies is being developed and tested using AI. For example, new materials development rests on quantum mechanical equations of multi-atom

geometries that are easy to write down but intractable to solve in closed form: AI-enabled computer calculations are the only way these new materials – with fantastic weight, strength, thermal, electronic, and other properties – can be engineered. A U.S. military unmatched in its use of AI is therefore not only essential, but also key to all kinds of military innovation.

Another question I was frequently asked as Secretary of Defense is whether there will be two Internets, one U.S.-led and one China-led. There will, indeed, surely be two tech ecosystems. That is not a choice the United States can make; Xi Jinping has announced China's intention to make it so. Moreover, in geopolitical terms, China's development has not taken the path that Americans and their allies had naively hoped for as recently as a few years ago. China has not embraced values of universal valence, as America does, at least on its best days, but instead embraces values that are distinctly and exclusively ethnocentrically Chinese. Thus, the United States and China have become locked in a titanic geostrategic struggle incomparably more complex than that between the United States and Soviet Union during the Cold War. The two Cold War opponents did not trade with each other in high-tech goods. The United States and China do.

It is essential to any U.S. Secretary of Defense that America continue to be unsurpassed in all the emerging fields of technology, including, of course, in AI. Prevailing in the competition will require a new geostrategic playbook for competition with China with chapters on defense, offense, and new alliances. Defense encompasses carefully tailored restrictions on critical sensitive technology that could make its way into China. Far more important than tech defense to limit China is tech offense to improve America: robust federal research and development funding and an overall innovative climate – encompassing regulation, education and immigration, capital markets, and so on – that is maximally simulative of superiority in AI and other fields (all, as noted, enabled by AI). Finally, recalling that China makes up but one-half of Asia and one-fifth of the world, it is essential that the U.S.-led tech ecosystem embrace most of the rest of humanity. It is unlikely that China or other potential military opponents of the United States will respect the same moral scruples that the United States applies to itself. But this essay suggests that the United States will not be disadvantaged in such an asymmetrical competition since good engineering design can accommodate both high performance and good ethics. Assuming the United States retains its historic values and does not forget to apply them to AI and other new technologies in the manner described here, the result will be a peaceful and progressive world for most.

AUTHOR'S NOTE

I am grateful to D. J. Patil, former U.S. Chief Data Scientist, for helpful comments on this essay.

ABOUT THE AUTHOR

Ash Carter, a Fellow of the American Academy since 1992, is Director of the Belfer Center for Science and International Affairs and Belfer Professor of Technology and Global Affairs at Harvard Kennedy School, and a member of the President's Council of Advisors on Science and Technology. He served as U.S. Secretary of Defense from 2015–2017. He is the author of *Inside the Five-Sided Box: Lessons from a Lifetime of Leadership in the Pentagon* (2019) and *Preventive Defense: A New Security Strategy for America* (with William J. Perry, 1997) and editor of *Keeping the Edge: Managing Defense for the Future* (with John P. White, 2001).

ENDNOTES

[1] The term AI has been around a long time and, for our purposes, means all kinds of advanced techniques: machine learning, neural networks, deep learning, and even just "big data." It does not make the distinction of "artificial general intelligence" (AGI) since the definition and meaning of AGI are not precise, and its "singularity" date–when AI matches or surpasses human intelligence–is elusive. The real singularity in the existence of technology will be when we can achieve human immortality, either digital or biological. "Immortality" might even happen before AGI.

[2] What about "full disclosure," "opt in/opt out," "anonymity," "it is impossible with such complicated systems"? All these are much more dubious, as we shall see.

[3] U.S. Department of Defense, "Autonomy in Weapon Systems," Directive No. 3000.09, November 21, 2012.

Distrust of Artificial Intelligence: Sources & Responses from Computer Science & Law

Cynthia Dwork & Martha Minow

Social distrust of AI stems in part from incomplete and faulty data sources, inappropriate redeployment of data, and frequently exposed errors that reflect and amplify existing social cleavages and failures, such as racial and gender biases. Other sources of distrust include the lack of "ground truth" against which to measure the results of learned algorithms, divergence of interests between those affected and those designing the tools, invasion of individual privacy, and the inapplicability of measures such as transparency and participation that build trust in other institutions. Needed steps to increase trust in AI systems include involvement of broader and diverse stakeholders in decisions around selection of uses, data, and predictors; investment in methods of recourse for errors and bias commensurate with the risks of errors and bias; and regulation prompting competition for trust.

Works of imagination, from *Frankenstein* (1818) to the film *2001: A Space Odyssey* (1968) and the *Matrix* series (1999–2021), explore fears that human-created artificial intelligences threaten human beings due to amoral logic, malfunctioning, or the capacity to dominate.[1] As computer science expands from human-designed programs spelling out each step of reasoning to programs that automatically learn from historical data, infer outcomes for individuals not yet seen, and influence practices in core areas of society – including health care, education, transportation, finance, social media, retail consumer businesses, and legal and social welfare bureaucracies – journalistic and scholarly accounts have raised questions about reliability and fairness.[2] Incomplete and faulty data sources, inappropriate redeployment of data, and frequently exposed errors amplify existing social dominance and cleavages. Add mission creep – like the use of digital tools intended to identify detainees needing extra supports upon release to instead determine release decisions – and it is no wonder that big data and algorithmic tools trigger concerns over loss of control and spur decay in social trust essential for democratic governance and workable relationships in general.[3]

© 2022 by Cynthia Dwork & Martha Minow
Published under a Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) license
https://doi.org/10.1162/DAED_a_01918

Failures to name and comprehend the basic terms and processes of AI add to specific sources of distrust. Examining those sources, this essay ends with potential steps forward, anticipating both short-term and longer-term challenges.

Artificial intelligence signifies a variety of technologies and tools that can solve tasks requiring "perception, cognition, planning, learning, communication, and physical actions," often learning and acting without oversight by their human creators or other people.[4] These technologies are already much used to distribute goods and benefits by governments, private companies, and other private actors.

Trust means belief in the reliability or truth of a person or thing.[5] Associated with comfort, security, and confidence, its absence infers doubt about the reliability or truthfulness of a person or thing. That doubt generates anxieties, alters behaviors, and undermines cooperation needed for private and public action. *Distrust* is corrosive.

Distrust is manifested in growing calls for regulation, the emergence of watchdog and lobbying groups, and the explicit recognition of new risks requiring monitoring by corporate audit committees and accounting firms.[6] Critics and advocates alike acknowledge that increasing deployment of AI could have unintended but severe consequences for human lives, ranging from impairments of friendships to social disorder and war.[7] These concerns multiply in a context of declining trust in government and key private institutions.[8]

An obvious source of distrust is evidence of unreliability. Unreliability could arise around a specific task, such as finding that your child did not run the errand to buy milk as you requested. Or it could follow self-dealing: did your child keep the change from funds used to purchase the milk rather than returning the unused money to you? Trust is needed when we lack the time or ability to oversee each task to ensure truthful and accurate performance and devotion to the interests of those relying on the tasks being done.

Political theorist Russell Hardin explains trust as "encapsulated interest, in which the truster's expectations of the trusted's behavior depend on assessments of certain motivations of the trusted. I trust you because your interests encapsulate mine to some extent – in particular, because you want our relationship to continue."[9] Trust accordingly is grounded in the truster's assessment of the intentions of the trusted with respect to some action.[10] Trust is strengthened when I believe it is in your interest to adhere to my interests in the relevant matter.[11] Those who rely on institutions, such as the law, have reasons to believe that they comport with governing norms and practices rather than serving other interests.

Trust in hospitals and schools depends on assessments of the reliability of the institution and its practices in doing what it promises to do, as well as its responses to inevitable mistakes.[12] With repeated transactions, trust depends not only

on results, but also on discernable practices reducing risks of harm and deviation from expected tasks. Evidence that the institution or its staff serves the interests of intended beneficiaries must include guards against violation of those interests. Trust can grow when a hospital visibly uses good practices with good results and communicates the measures to minimize risks of bad results and departures from good practices.

External indicators, such as accreditation by expert review boards, can signal adherence to good practices and reason to expect good results. External indicators can come from regulators who set and enforce rules, such as prohibitions of self-dealing through bans on charging more than is justifiable for a procedure and prohibiting personal or institutional financial interests that are keyed to the volume of referrals or uses.[13] Private or governmental external monitors must be able to audit the behavior of institutions.[14] External review will not promote trust if external monitors are not themselves trusted. In fact, disclosure amid distrust can feed misunderstandings.[15]

Past betrayals undermine trust. Personal and collective experiences with discrimination or degradation – along lines of race, class, gender, or other personal characteristics – especially create reasons for suspicion if not outright distrust. Similarly, experiences with self-interested companies that make exploitative profits can create or sustain distrust. Distrust and the vigilance it inspires may itself protect against exploitation.[16]

These and further sources of distrust come with uses of AI, by which we mean: a variety of techniques to discern patterns in historical "training" data that are determinative of status (is the tumor benign?) or predictive of a future outcome (what is the likelihood the student will graduate within four years?). The hope is that the patterns discerned in the training data will extend to future unseen examples. Algorithms trained on data are "learned algorithms." These learned algorithms classify and score individuals as the system designer chose, equitably or not, to represent them to the algorithm. These representations of individuals and "outcomes" can be poor proxies for the events of interest, such as using re-arrest as a proxy for recidivism or a call to child protective services as a proxy for child endangerment.[17]

Distrust also results from the apparent indifference of AI systems. Learned algorithms lack indications of adherence to the interests of those affected by their use. They also lack apparent conformity with norms or practices legible to those outside of their creation and operations.

When designed solely at the directive of governments and companies, AI may only serve the interests of governments and companies – and risk impairing the interests of others.

Despite sophisticated techniques to teach algorithms from data sets, *there is no ground truth* available to check whether the results match reality. This is a basic challenge for ensuring reliable AI. We can prove that the

learned algorithm is indeed the result of applying a specific learning technique to the training data, but when the learned algorithm is applied to a previously unseen individual, one not in the training data, we do not have proof that the outcome is correct in terms of an underlying factual basis, rather than inferences from indirect or arbitrary factors. Consider an algorithm asked to predict whether a given student will graduate within four years. This is a question about the future: when the algorithm is applied to the data representing the student, the answer has not yet been determined. A similar quandary surrounds risk scoring: what is the "probability" that an individual will be re-arrested within two years? This question struggles to make sense even mathematically: what is the meaning of the "probability" of a nonrepeatable event?[18] Is what we perceive as randomness in fact certainty, if only we had sufficient contextual information and computing power? Inferences about the future when predicated on limited or faulty information may create an illusion of truth, but illusion it is.

Further problems arise because techniques for building trust are too often unavailable with algorithms used for scoring and categorizing people for public or private purposes. Familiar trust-building techniques include transparency so others can see inputs and outcomes, opportunities for those affected to participate in designing and evaluating a system and in questioning its individual applications, monitoring and evaluation by independent experts, and regulation and oversight by government bodies.

Trust in the fairness of legal systems increases when those affected participate with substantive, empowering choices within individual trials or panels reviewing the conduct of police and other officials. Could participation of those affected by AI help build trust in uses of AI?[19] Quite apart from influencing outcomes, participation gives people a sense that they are valued, heard, and respected.[20] Participatory procedures signal fairness, help to resolve uncertainties, and support deference to results.[21] Following prescribed patterns also contributes to the perceived legitimacy of a dispute resolution system.[22]

But there are few if any roles for consumers, criminal defendants, parents, or social media users to raise questions about the algorithms used to guide the allocation of benefits and burdens. Nor are there roles for them in the construction of the information-categorizing algorithms. Opportunities to participate are not built into the design of algorithms, data selection and collection protocols, or the testing, revision, and use of learning algorithms. Ensuring a role for human beings to check algorithmic processes can even be a new source of further inaccuracies. An experiment allowing people to give feedback to an algorithmically powered system actually showed that participation lowered trust – perhaps by exposing people to the scope of the system's inaccuracies.[23]

Suggestions for addressing distrust revolve around calls for "explainability" and ensuring independent entities access to the learned algorithms themselves.[24]

"Access" can mean seeing the code, examining the algorithm's outputs, and reviewing the choice of representation, sources of training data, and demographic benchmarking.[25] But disclosure of learning algorithms themselves has limited usefulness in the absence of data features with comprehensible meanings and explanations of weight determining the contribution of each feature to outcomes. Machine learning algorithms use mathematical operations to generate data features that almost always are not humanly understandable, even if disclosed, and whose learned combinations would do nothing to explain outcomes, even to expert auditors.

Regulation can demand access and judgments by qualified experts and, perhaps more important, require behavior attentive not only to narrow interests but also to broader public concerns. Social distrust of X-rays produced demands for regulation; with regulation, professional training, and standards alert to health effects, X-rays gained widespread trust.[26] Yet government regulators and independent bodies can stoke public fears if they contribute to misinformation and exaggerate risks.[27]

For many, reliance on AI arouses fears of occupational displacement. Now white collar as well as blue collar jobs seem at risk. One study from the United Kingdom reported that more than 60 percent of people surveyed worry that their jobs will be replaced by AI. Many believe that their jobs and opportunities for their children will be disrupted.[28] More than one-third of young Americans report fears about technology eliminating jobs.[29] Despite some predictions of expanded and less-repetitive employment, no one can yet resolve doubts about the future.[30] Foreboding may be exacerbated by awareness that, by our uses of technology, we contribute to the trends we fear. Many people feel forced to use systems such as LinkedIn or Facebook.[31] People report distrust of the Internet but continue to increase their use of it.[32]

Some distrust AI precisely because human beings are not visibly involved in decisions that matter to human beings. Yet even the chance to appeal to a human is insufficient when individuals are unaware that there is a decision or score affecting their lives.

As companies and governments increase their use of AI, distrust mounts considerably with misalignment of interests. Airbnb raised concerns when it acquired Trooly Inc., including its patented "trait analyzer" that operates by scouring blogs, social networks, and commercial and public databases to derive personality traits. The patent claims that "the system determines a trustworthiness score of the person based on the behavior and personality trait metrics using a machine learning system," with the weight of each personality trait either hard coded or inferred by a machine learning model.[33] It claims to identify

traits as "badness, anti-social tendencies, goodness, conscientiousness, openness, extraversion, agreeableness, neuroticism, narcissism, Machiavellianism, or psychopathy."[34] Although Airbnb asserts that the company is not currently deploying this software,[35] the very acquisition of a "trait analyzer" raises concerns that the company refuses to encapsulate the interests of those affected.[36]

Examples of practices harming and contrary to the interests of users abound in social media platforms, especially around demonstrated biases and invasions of privacy. Although social media companies offer many services that appeal to users, the companies have interests that diverge systematically from those of users. Platform companies largely profit off data generated by each person's activities on the site. Hence, the companies seek to maximize user "engagement." Each new data point comes when a user does – or does not – click on a link or hit a "like" button. The platform uses that information to tailor content for users and to sell their information to third parties for targeted advertising and other messages.[37] Chamath Palihapitiya, former vice president for "user growth" for Facebook, has claimed that Facebook is addictive by design.[38] Sean Parker, an original Facebook investor, has acknowledged that the site's "like" button and news feed keep users hooked by exploiting people's neurochemical vulnerabilities.[39]

Privacy loss is a particular harm resented by many. Privacy can mean seclusion, hiding one's self, identity, and information; it can convey control over one's personal information and who can see it; it can signal control over sensitive or personal decisions, without interference from others; or it can mean protection against discrimination by others based on information about oneself. All these meanings matter in the case of Tim Stobierski, who, shortly after starting a new job at a publishing house, was demonstrating a Facebook feature to his boss when an advertisement for a gay cruise appeared on his news feed.[40] He wondered, "how did Facebook know that I was interested in men, when I had never told another living soul, and when I certainly had never told Facebook?"[41] The Pew Research Center showed that about half of all Facebook users feel discomfort about the site's collection of their interests, while 74 percent of Facebook users did not know how to find out how Facebook categorized their interests or even how to locate a page listing "your ad preferences."[42] A platform's assumptions remain opaque even as users resent the loss of control over their information and the secret surveillance.[43]

Tech companies may respond that users can always quit. Here, too, a conflict of interests is present. Facebook exposes individuals to psychological manipulation and data breaches to degrees that they cannot imagine.[44] Most users do not even know how Facebook uses their data or what negative effects can ensue.[45] The loss of control compounds the unintended spread of personal information.

The interests of tech platforms and users diverge further over hateful speech. Facebook's financial incentive is to keep or even elicit outrageous posts because they attract engagement (even as disagreement or disgust) and hence produce

additional monetizable data points.[46] Facebook instructs users to hide posts they do not like, or to unfollow the page or person who posted it, and, only as a third option, to report the post to request its removal.[47] Under pressure, Facebook established an oversight review board and charged it with evaluating (only an infinitesimal fraction of) removal decisions. Facebook itself determines which matters proceed to review.[48] Directed to promote freedom of speech, not to guard against hatred or misinformation, the board has so far done little to guard against fomented hatred and violence.[49]

Large tech companies are gatekeepers; they can use their position and their knowledge of users to benefit their own company over others, including third parties that pay for their services.[50] As one observer put it, "social media is cloaked in this language of liberation while the corporate sponsors (Facebook, Google *et al*.) are progressing towards ever more refined and effective means of manipulating individual behavior (behavioral targeting of ads, recommendation systems, reputation management systems etc.)."[51]

The processes of AI baffle the open and rational debates supporting democracies, markets, and science that have existed since the Enlightenment. AI practices can nudge and change what people want, know, and value.[52] Differently organized, learned algorithms could offer people some control over site architecture and content moderation.[53]

Dangers from social media manipulation came to fruition with the 2020 U.S. presidential election. Some conventional media presented rumors and falsehood, but social media initiated and encouraged misinformation and disinformation, and amplified their spread, culminating in the sweeping erroneous belief that Donald Trump rather than Joe Biden had won the popular vote. False claims of rigged voting machines, despite the certification of state elections, reflected and inflamed social distrust.[54] The sustainability of our democratic governance systems is far from assured.

Building trust around AI can draw on commitments to participation, useable explanations, and iterative improvements. Hence, people making and deploying AI should involve broader and diverse stakeholders in decisions around what uses algorithms are put to; what data, with which features, are used to train the algorithms; what criteria are used in the training process to evaluate classifications or predictions; and what methods of recourse are available for raising concerns about and securing genuine responsive action to potentially unjust methods or outcomes. Creative and talented people have devised AI algorithms able to infer our personal shopping preferences; they could deploy their skills going forward to devise opportunities for those affected to participate in identifying gaps and distortions in data. Independent experts in academic and nonprofit settings – if given access to critical information – could provide much-needed audits of algo-

rithmic applications and assess the reliability and failures of the factors used to draw inferences.

Investment in participatory and information-sharing efforts should be commensurate with the risks of harms. Otherwise, the risks are entirely shifted to the consumers, citizens, and clients who are subjected to the commercial and governmental systems that deploy AI algorithms.

As AI escalates, so should accessible methods of recourse and correction. Concerns for people harmed by harassment on social media; biased considerations in employment, child protection, and other governmental decisions; and facial recognition technologies that jeopardize personal privacy and liberty will be echoed by known and unknown harms in finance, law, health care, policing, and war-making. Software systems to enable review and to redress mistakes should be built, and built to be meaningful. Designers responding that doing so would be too expensive or too difficult given the scale enabled by the use of AI algorithms are scaling irresponsibly. Responsible scaling demands investment in methods of recourse for errors and bias commensurate with the risks of errors and bias. AI can and must be part of the answer in addressing the problems created by AI, but so must strengthened roles for human participation. Government by the consent of the governed needs no less.[55]

Self-regulation and self-certification, monitoring by external industry and consumer groups, and regulation by government can tackle misalignment and even clashes in the interests of those designing the learning algorithms and those affected by them. Entities should compete in the marketplace for trust and reputation, face ratings by external monitors, and contribute to the development of industry standards. Trust must be earned.

AUTHORS' NOTE

We thank Christian Lansang, Maroussia Lévesque, and Serena Wong for thoughtful research and advice for this essay.

ABOUT THE AUTHORS

Cynthia Dwork, a Fellow of the American Academy since 2008, is the Gordon McKay Professor of Computer Science in the John Paulson School of Engineering and Applied Sciences and Radcliffe Alumnae Professor in the Radcliffe Institute for Advanced Study at Harvard University. Her work has established the pillars of fault-tolerant distributed systems, modernized cryptography to the ungoverned

interactions of the Internet and the era of quantum computing, revolutionized privacy-preserving statistical data analysis, and launched the field of algorithmic fairness.

Martha Minow, a Fellow of the American Academy since 1992, is the 300th Anniversary University Professor and Distinguished Service Professor at Harvard University. She also serves as Cochair of the American Academy's project on Making Justice Accessible: Designing Legal Services for the 21st Century. She is the author of, most recently, *Saving the News: Why the Constitution Calls for the Government to Act to Preserve the Freedom of Speech* (2021), *When Should Law Forgive?* (2019), and *In Brown's Wake: Legacies of America's Educational Landmark* (2010).

ENDNOTES

[1] See "AI in Pop Culture," ThinkAutomation, https://www.thinkautomation.com/bots-and-ai/ai-in-pop-culture/.

[2] Darrell M. West and John. R. Allen, *Turning Point: Policymaking in the Era of Artificial Intelligence* (Washington, D.C.: Brookings Institution Press, 2020).

[3] Although many are optimistic about new technologies, concerns over loss of control are growing. See Ethan Fast and Eric Horvitz, "Long-Term Trends in the Public Perception of Artificial Intelligence," in *AAAI '17: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence* (Menlo Park: Calif.: Association for the Advancement of Artificial Intelligence, 1979), 963.

[4] National Security Commission on Artificial Intelligence, *Final Report: Artificial Intelligence in Context* (Washington, D.C.: National Security Commission on Artificial Intelligence, 2021), https://reports.nscai.gov/final-report/ai-in-context/.

[5] "Trust," Oxford English Dictionary Online, http://www.oed.com/view/Entry/207004 (accessed November 16, 2020).

[6] See Deloitte, *Managing Algorithmic Risks: Safeguarding the Use of Complex Algorithms and Machine Learning* (London: Deloitte Development, LLC, 2017), https://www2.deloitte.com/content/dam/Deloitte/us/Documents/risk/us-risk-algorithmic-machine-learning-risk-management.pdf; Simson L. Garfinkel, "A Peek at Proprietary Algorithms," *American Scientist* 105 (6) (2017), https://www.americanscientist.org/article/a-peek-at-proprietary-algorithms; and Stacy Meichtry and Noemie Bisserbe, "France's Macron Calls for Regulation of Social Media to Stem 'Threat to Democracy,'" *The Wall Street Journal*, January 29, 2021, https://www.wsj.com/articles/frances-macron-calls-for-regulation-of-social-media-to-stem-threat-to-democracy-11611955040.

[7] Such as National Security Commission on Artificial Intelligence, *Final Report*, "Chapter 7: Establishing Justified Confidence in AI Systems" and "Chapter 8: Upholding Democratic Values: Privacy, Civil Liberties, and Civil Rights in Uses of AI for National Security." See also Executive Office of the President, *Big Data: Seizing Opportunities, Preserving Values* (Washington, D.C.: Executive Office of the President, 2014), https://obamawhitehouse.archives.gov/sites/default/files/docs/big_data_privacy_report_may_1_2014.pdf.

[8] Lee Rainie, Scott Keeter, and Andrew Perrin, "Trust and Distrust in America," Pew Research Center, https://www.pewresearch.org/politics/2019/07/22/trust-and-distrust-in-america/. On the psychology of distrust, see Roy J. Lewicki and Edward C. Tomlinson, "Distrust," Beyond Intractability, December 2003, https://www.beyondintractability.org/essay/distrust. See also Embedded Ethics @ Harvard, https://embeddedethics.seas.harvard.edu/; "Ethics, Computer, and AI: Perspectives from MIT, Human Contexts and Ethics," MIT News, March 18, 2019, https://news.mit.edu/2019/ethics-computing-and-ai-perspectives-mit-0318; and Berkeley Computing, Data Science, and Society, https://data.berkeley.edu/hce.

[9] Russell Hardin, *Trust and Trustworthiness* (New York: Russell Sage Foundation, 2002), xix.

[10] Ibid., xx.

[11] Ibid., 4.

[12] See Pierre Lauret, "Why (and How to) Trust Institutions? Hospitals, Schools, and Liberal Trust," *Rivista di Estetica* 68 (2018): 41–68, https://doi.org/10.4000/estetica.3455.

[13] AMA Council on Ethical and Judicial Affairs, "AMA Code of Medical Ethics' Opinions on Physicians' Financial Interests," Opinion 8.0321–Physicians' Self-Referral, *AMA Journal of Ethics* (August 2015), https://journalofethics.ama-assn.org/article/ama-code-medical-ethics-opinions-physicians-financial-interests/2015-08. For analogous treatment of AI, see Matthew Hutson, "Who Should Stop Unethical A.I.?" *The New Yorker*, February 15, 2021, https://www.newyorker.com/tech/annals-of-technology/who-should-stop-unethical-ai.

[14] See Michael Kearns and Aaron Roth, "Ethical Algorithm Design Should Guide Technology Regulation," The Brookings Institution, January 13, 2020, https://www.brookings.edu/research/ethical-algorithm-design-should-guide-technology-regulation/.

[15] See Ethan Zuckerman, *Distrust: Why Losing Faith in Institutions Provides the Tools to Transform Them* (New York: W. W. Norton & Company, 2021), 60 (describing tendencies of people living with broken institutions to wrongly see patterns and conspiracies in random occurrences).

[16] Roderick M. Kramer, "Rethinking Trust," *Harvard Business Review*, June 2009, https://hbr.org/2009/06/rethinking-trust; and Christopher B. Yenkey, "The Outsider's Advantage: Distrust as a Deterrent to Exploitation," *American Journal of Sociology* 124 (3) (2018): 613.

[17] Virginia Eubanks, *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor* (New York: St. Martin's Press, 2018).

[18] See, for example, A. Philip Dawid, "On Individual Risk," *Synthese* 194 (2017); and Cynthia Dwork, Michael P. Kim, Omer Reingold, et al., "Outcome Indistinguishability," in *STOC 2021: Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing* (New York: Association for Computing Machinery, 2021). For a treatment of individual probabilities in risk assessment tools, see Peter B. Imrey and A. Philip Dawid, "A Commentary on the Statistical Assessment of Violence and Recidivism Risks," *Statistics and Public Policy* 2 (1) (2015): 1–18; and Kristian Lum, David B. Dunson, and James E. Johndrow, "Closer than They Appear: A Bayesian Perspective on Individual-Level Heterogeneity in Risk Assessment," arXiv (2021), https://arxiv.org/abs/2102.01135.

[19] Tom R. Tyler, "Procedural Justice, Legitimacy, and the Effective Rule of Law," *Crime and Justice* 30 (2003): 283, https://www.jstor.org/stable/1147701?seq=1#metadata_info_tab_contents.

[20] See E. Allan Lind and Tom R. Tyler, eds., *The Social Psychology of Procedural Justice* (New York: Springer, 1988).

[21] Kees van den Bos, Lynn van der Velden, and E. Allan Lind, "On the Role of Perceived Procedural Justice in Citizens' Reactions to Government Decisions and the Handling of Conflicts," *Utrecht Law Review* 10 (4) (2014): 1–26, https://www.utrechtlawreview.org/articles/abstract/10.18352/ulr.287/.

[22] Rebecca Hollander-Blumoff and Tom R. Tyler, "Procedural Justice and the Rule of Law: Fostering Legitimacy in Alternative Dispute Resolution," *Journal of Dispute Resolution* 1 (2011).

[23] Donald R. Honeycutt, Mahsan Nourani, and Eric D. Ragan, "Soliciting Human-in-the-Loop User Feedback for Interactive Machine Learning Reduces User Trust and Impressions of Model Accuracy," in *Proceedings of the Eighth AAAI Conference on Human Computation and Crowdsourcing* (Menlo Park, Calif.: Association for the Advancement of Artificial Intelligence, 2020), 63–72, https://ojs.aaai.org/index.php/HCOMP/article/view/7464.

[24] See, for example, David Leslie, "Project ExplAIn," The Alan Turing Institute, December 2, 2019, https://www.turing.ac.uk/news/project-explain, which describes six kinds of "explanation types," including identifying who is involved in the development and management of an AI system and whom to contact for human review, as well as the effects that the AI system has on an individual and on wider society.

[25] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, et al., "Datasheets for Datasets," arXiv (2018), https://arxiv.org/abs/1803.09010; and Margaret Mitchell, Simone Wu, Andrew Zaldivar, et al., "Model Cards for Model Reporting," in *FAT* '19: Proceedings of the Conference on Fairness, Accountability, and Transparency* (New York: Association for Computing Machinery, 2019), 220–229. See also Cynthia Dwork, Michael P. Kim, Omer Reingold, et al., "Outcome Indistinguishability," in STOC *2021: Proceedings of the 53rd Annual* ACM SIGACT *Symposium on Theory of Computing* (New York: Association for Computing Machinery, 2021).

[26] See Antony Denman, S. Parkinson, and Christopher John Groves-Kirby, "A Comparative Study of Public Perception of Risks from a Variety of Radiation and Societal Risks," presented at the 11th International Congress of the International Radiation Protection Association, Madrid, Spain, May 23–28, 2004.

[27] John M. Osepchuk, "A History of Microwave Heating Applications," *IEEE Transactions on Microwave Theory and Techniques* 32 (9) (1984): 1200, 1213.

[28] Jacob Douglas, "These American Workers Are the Most Afraid of A.I. Taking Their Jobs," CNBC, November 7, 2019 (37 percent of people surveyed aged eighteen to twenty-four expressed fear AI will take their jobs), https://www.cnbc.com/2019/11/07/these-american-workers-are-the-most-afraid-of-ai-taking-their-jobs.html; and "Technically Redundant: Six-in-10 Fear Losing Their Jobs to AI," Industry Europe, November 3, 2019, https://industryeurope.com/technically-redundant-six-in-10-fear-losing-their-jobs-to-ai/.

[29] Douglas, "These Americans Are the Most Afraid of AI Taking Their Jobs."

[30] James E. Bessen, Stephen Impink, Lydia Reichensperger, and Robert Seamans, "The Business of AI Startups" (Boston: Boston University School of Law, 2018), https://ssrn.com/abstract=3293275 or http://dx.doi.org/10.2139/ssrn.3293275.

[31] Since the writing of this essay, Facebook has been rebranded as Meta.

[32] Lee Raine and Janna Anderson, "Theme 3: Trust Will Not Grow, but Technology Usage Will Continue to Rise as a 'New Normal' Sets in," Pew Research Center, August 10, 2017, https://www.pewresearch.org/internet/2017/08/10/theme-3-trust-will-not-grow-but-technology-usage-will-continue-to-rise-as-a-new-normal-sets-in/.

[33] Sarabjit Singh Baveja, Anish Das Sarma, and Nilesh Dalvi, United States Patent No. 9070088 B1: Determining Trustworthiness and Compatibility of a Person, June 30, 2015, https://patentimages.storage.googleapis.com/36/36/7e/db298c5d3b280c/US9070088.pdf (accessed January 11, 2022).

[34] Ibid., 3–4.

[35] Aaron Holmes, "Airbnb Has Patented Software that Digs through Social Media to Root Out People Who Display 'Narcissism or Psychopathy,'" *Business Insider*, January 6, 2020, https://www.businessinsider.com/airbnb-software-predicts-if-guests-are-psychopaths-patent-2020-1.

[36] See text above from endnotes 14–17.

[37] Tero Karppi, *Disconnect: Facebook's Affective Bonds* (Minneapolis: University of Minnesota Press, 2018).

[38] Ibid.

[39] Tero Karppi and David B. Nieborg, "Facebook Confessions: Corporate Abdication and Silicon Valley Dystopianism," *New Media & Society* 23 (9) (2021); Sarah Friedman, "How Your Brain Responds Every Time Your Insta Post Gets a 'Like,'" *Bustle*, September 21, 2019, https://www.bustle.com/p/how-your-brain-responds-to-a-like-online-shows-the-power-of-social-media-18725823; and Trevor Haynes, "Dopamine, Smartphones, and You: A Battle for Your Time," Science in the News Blog, Harvard Medical School, May 1, 2018, https://sitn.hms.harvard.edu/flash/2018/dopamine-smartphones-battle-time/.

[40] Tim Stobierski, "Facebook Ads Outed Me," INTO, May 3, 2018, https://www.intomore.com/you/facebook-ads-outed-me/#.

[41] Ibid.

[42] Paul Hitlin and Lee Rainie, "Facebook Algorithms and Personal Data," Pew Research Center, January 16, 2019, https://www.pewresearch.org/internet/2019/01/16/facebook-algorithms-and-personal-data/. New "post-cookie" advertising schemes enlist the browser to perform user categorization previously carried out by advertising networks. Heralded as privacy-preserving because the browser is local to the user's machine, these systems are designed to carry out the same segmentation that so many find objectionable; see Bennet Cyphers, "Google's FLoC Is a Terrible Idea," March 3, 2021, Electronic Frontier Foundation. Standard notions of privacy do not ensure fair treatment. Cynthia Dwork and Deirdre K. Mulligan, "It's Not Privacy and It's Not Fair," *Stanford Law Review* 66 (2013), https://www.stanfordlawreview.org/online/privacy-and-big-data-its-not-privacy-and-its-not-fair/ ("The urge to classify is human. The lever of big data, however, brings ubiquitous classification, demanding greater attention to the values embedded and reflected in classifications, and the roles they play in shaping public and private life.")

[43] Moreover, "privacy solutions can hinder efforts to identify classifications that unintentionally produce objectionable outcomes–for example, differential treatment that tracks race or gender–by limiting the availability of data about such attributes." Dwork and Mulligan, "It's Not Privacy and It's Not Fair."

[44] See Karppi and Nieborg, "Facebook Confessions," 11.

[45] Ibid.

[46] Karppi, *Disconnect: Facebook's Affective Bonds*.

[47] Eugenia Siapera and Paloma Viejo-Otero, "Governing Hate: Facebook and Digital Racism," *Television & New Media* 22 (2) (2020): 112–113, 122.

[48] Evelyn Douek, "What Kind of an Oversight Board Have You Given Us?" *University of Chicago Law Review Online*, May 11, 2020, https://lawreviewblog.uchicago.edu/2020/05/11/fb-oversight-board-edouek/.

[49] See Andrew Marantz, "Why Facebook Can't Fix Itself," *The New Yorker*, October 12, 2020.

[50] Lina M. Khan, "Sources of Tech Platform Power," *Georgetown Law Technology Review* 2 (2) (2018): 325.

[51] Joshua-Michéle Ross, "The Question Concerning Social Technology," Radar, May 18, 2009, http://radar.oreilly.com/2009/05/the-question-concerning-social.html; and "Do Social Media Threaten Democracy?" *The Economist*, November 4, 2017, https://www.economist.com/leaders/2017/11/04/do-social-media-threaten-democracy.

[52] Ibid.; and *The Social Dilemma*, dir. Jeff Orlowski (Boulder, Colo.: Exposure Labs, Argent Pictures, The Space Program, 2020).

[53] See Daphne Keller, "The Future of Platform Power: Making Middleware Work," *Journal of Democracy* 32 (3) (2021), https://www.journalofdemocracy.org/articles/the-future-of-platform-power-making-middleware-work/.

[54] See, for example, Aaron Blake, "Trump's 'Big Lie' Was Bigger Than Just a Stolen Election," *The Washington Post*, February 12, 2021, https://www.washingtonpost.com/politics/2021/02/12/trumps-big-lie-was-bigger-than-just-stolen-election/; Melissa Block, "Can The Forces Unleashed By Trump's Big Election Lie Be Undone?" NPR, January 16, 2021, https://www.npr.org/2021/01/16/957291939/can-the-forces-unleashed-by-trumps-big-election-lie-be-undone; and Christopher Giles and Jake Horton, "U.S. Election 2020: Is Trump Right about Dominion Machines?" BBC, November 17, 2020, https://www.bbc.com/news/election-us-2020-54959962.

[55] See Richard Fontaine and Kara Frederick, "Democracy's Digital Defenses," *The Wall Street Journal*, May 7, 2021, https://www.wsj.com/amp/articles/democracys-digital-defenses-11620403161.

Democracy & Distrust in an Era of Artificial Intelligence

Sonia K. Katyal

Our legal system has historically operated under the general view that courts should defer to the legislature. There is one significant exception to this view: cases in which it appears that the political process has failed to recognize the rights or interests of minorities. This basic approach provides much of the foundational justifications for the role of judicial review in protecting minorities from discrimination by the legislature. Today, the rise of AI decision-making poses a similar challenge to democracy's basic framework. As I argue in this essay, the rise of three trends – privatization, prediction, and automation in AI – have combined to pose similar risks to minorities. In this essay, I outline what a theory of judicial review would look like in an era of artificial intelligence, analyzing both the limitations and the possibilities of judicial review of AI. Here, I draw on cases in which AI decision-making has been challenged in courts, to show how concepts of due process and equal protection can be recuperated in a modern AI era, and even integrated into AI, to provide for better oversight and accountability.

Almost forty years ago, in an elegant essay published in *Dædalus*, J. David Bolter wrote, "artificial intelligence is compelling and controversial, not for its practical achievements, but rather for the metaphor that lies behind the programs: the idea that human beings should be seen as nature's digital computers."[1] "The computer," Bolter continued, "is a mirror of human nature, just as any invention reflects to some extent the intellect and character of its inventor. But it is not a perfect mirror; it affects and perhaps distorts our gaze, magnifying certain human capacities . . . and diminishing others."[2]

As the author points out, a study of AI, which intrinsically compels us to compare mind and machine, reveals the distortions and inaccuracies within each realm. Metaphor, in these contexts, can be a useful way to parse the limits of comparison between humankind and machines. On this point, Bolter wrote, "we do not have to become religious converts to artificial intelligence in order to appreciate the computer metaphor. . . . Instead, we can ask in what ways the metaphor is apt and in what ways it may fail."[3] In other words, the study of artificial intelligence forces us to examine deep, compositional questions: What makes a hu-

© 2022 by Sonia K. Katyal
Published under a Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) license
https://doi.org/10.1162/DAED_a_01919

man? What makes a machine? And, most important, what makes something artificial, or intelligent?

To some extent, a similar set of compositional comparisons can be posed toward the relationship between law and democracy. Law is a metaphor of sorts – a set of artificial principles – that help us to move toward an ideal society; but the execution of law intrinsically requires us to compare the artifice of these ideals with the unpredictable reality of humanity and governance, thus revealing the distortions and inaccuracies within each realm. Just as computers function as imperfect mirrors of human nature – magnifying certain human capacities and diminishing others – law, too, is a reflection of these limitations and possibilities. And over time, the law has developed its own form of self-regulation to address these issues, stemming from the risks surrounding human fallibility. Our legal system has developed an architectural design of separate institutions, a system of checks and balances, and a vibrant tradition of judicial review and independence. Taken together, these elements compose part of the design of democracy.

Similar elements, I argue in this essay, must be part of the future of artificial intelligence. That is precisely why a study of AI is necessarily incomplete without addressing the ways in which regulation can play a role in improving AI accountability and governance. The issues surrounding algorithmic accountability demonstrate a deeper, more structural tension within a new generation of disputes regarding law and technology, and the contrast between public and private accountability. At the core of these issues, of course, lies the issue of trust: trust in AI, trust in humanity, and trust in the rule of law and governance. Here, the true potential of AI does not lie in the information we reveal to one another, but rather in the issues it raises about the interaction of technology, public trust, and the rule of law.

The rise of AI in decision-making poses a foundational challenge to democracy's basic framework. To recuperate trust in AI for humanity's sake, it is essential to employ design systems that integrate principles of judicial review as a foundational part of AI-driven architecture. My approach in this essay sketches out three dimensions: descriptive, analytic, and normative. First, I describe the background theory of judicial review to introduce a few themes that are relevant to exploring the intersection between AI and our legal system. Then I argue that a system of judicial review is especially needed in light of the rise of three trends that have fundamentally altered the course of AI decision-making: privatization (the increased role of private contractors in making governmental decisions); prediction (the increased focus on using AI to predict human behaviors, in areas as wide-ranging as criminal justice and marketing); and an increased reliance on automated decision-making. These three trends, I argue, have combined to create a perfect storm of conflict that calls into question the role of courts and regulation altogether, potentially widening the gap of protection for minorities in a world that will become increasingly reliant on AI.

Finally, I turn to the normative possibilities posed by these challenges. How can we ensure that software designers, drawn by traditional approaches to statistical, predictive analytics, are mindful of the importance of avoiding disparate treatment? What protections exist to ensure a potential road map for regulatory intervention? Here, drawing on cases in which AI decision-making has been challenged in the courts, I sketch out some ways due process and equal protection can be recuperated in a modern AI era, and even integrated into AI, to provide for better oversight and accountability.

The concept of judicial review, in the United States, has long drawn its force from a famous footnote – perhaps the most famous footnote ever written – in the 1938 case *U.S. vs. Carolene Products*, which involved a constitutional challenge to an economic regulation. In the opinion, written by Justice Harlan Stone, the Court drew a distinction between economic regulation and other kinds of legislation that might affect the interests of other groups. This distinction, buried in that "footnote four," transformed the law's approach to civil rights, underpinning the guarantee of equal protection under the Fourteenth Amendment for all citizens in the future.

For economic regulations, the opinion explained, courts should adopt a more deferential standard of review, erring on the side of trusting the legislature. However, when it was clear that a piece of legislation targeted "discrete and insular minorities," Justice Stone recommended employing a heightened standard of review and scrutiny over the legislation, demanding greater justification to defend its enaction.[4] "When prejudice against *discrete and insular minorities* may be a special condition," Stone wrote, "which tends seriously to curtail the operation of those political processes ordinarily to be relied upon to protect minorities," the law needs to exercise a more "searching inquiry" to justify its actions.

In the footnote, Justice Stone encapsulated a simple, elegant theory: we need the courts to safeguard minorities from regulations that might disregard or disadvantage their interests. Of course, this is not the only reason for why we need judicial review. The famed *Carolene* footnote later formed the backbone of a seminal book by John Hart Ely, *Democracy and Distrust: A Theory of Judicial Review*. Ely's work was essentially a longer explication of this idea: by integrating a healthy distrust of the political process, we can further safeguard democracy for the future. To say that the work is formative would be an understatement, as *Democracy and Distrust* has been described as "the single most cited work on constitutional law in the last century," and "a rite of passage" for legal scholars.[5] By developing the ideas embodied in Stone's footnote, Ely put forth a theory, known as "representation-reinforcement theory," which posits that courts should generally engage in a variety of situations, including cases in which it appears that the political process has failed to recognize the rights or interests of minorities, or where fundamental

rights are at stake. This basic theory provides much of the foundational thinking for justifying the role of the judiciary in protecting minorities from discrimination and charting a course for judicial review.

Ely's work has been interpreted to offer a vision of democracy as a function of procedural values, rather than substantive ones, by focusing on the way that judicial systems can create the conditions for a fair political process.[6] One example of this sort of process malfunction, Ely described, involved an intentional kind of disenfranchisement: "the ins," he observed, "are choking off the channels of political change to ensure that they will stay in and the outs will stay out."[7] A second kind of malfunction involved situations in which "no one is actually denied a voice or a vote," but representatives of a majority still systematically disadvantage minority interests "out of a simple hostility or prejudiced refusal to recognize commonalities of interest, and thereby denying that minority the protection afforded other groups by a representative system."[8]

Judicial review, under this approach, also exhorts us to explore whether particular groups face an undue constraint on their opportunity to participate in the political process.[9] For example, if minorities (or other groups) are constrained from participating fully in the political process, then the theory of representation-reinforcement focuses on proxy participation as a solution. Here, Ely reasoned, judges might stand in the place of minorities to ascertain the impact that they may face and take on the responsibility to craft a more inclusive solution. Or if fundamental rights are under threat, the Court should also intervene in order to preserve the integrity of the political process.

This basic theory undergirds much of the institutional and legal relationships between constitutional entitlements and the role of judges in this process. Like any other theory, Ely's approach is not perfect: it has been criticized, and rightfully so, for focusing too much on process at the expense of substantive constitutional rights.[10] But this theory of judicial review also yields both descriptive and normative insights into the government regulation of AI.

Reading Stone's and Ely's concerns in today's era of AI, one is immediately struck by their similarity of context. Both were concerned with the risk of majoritarian control, and designed systems of judicial review to actively protect minority interests. Today, those same concerns are almost perfectly replicated by certain AI-driven systems, suggesting that here, too, judicial review may be similarly necessary. And, normatively, just as judicial review is prescribed as a partial solution to address these risks of majoritarian control in a constitutional democracy, this insight holds similar limits and possibilities in the context of AI regulation.

Put another way, just as our political system often fails to represent the interests of demographic minorities, AI systems carry the same risks regarding the ab-

sence of representation and participation – but in private industry. Consider, for example, that one of the most central causes of biased outcomes in AI stems from an underlying problem of lack of representation among minority populations in the data sets used to train AI systems. Machine learning algorithms are, essentially, inherently regressive: they are trained on a body of data that is selected by designers or by past human practices. This process is the "learning" element in machine learning; the algorithm learns, for example, how to pair queries and results based on a body of data that produced satisfactory pairs in the past.[11] Thus, the quality of a machine learning algorithm's results often depends on the comprehensiveness and diversity of the data that it digests.[12]

As a result, bias in AI generally surfaces from these data-related issues of representation.[13] One problem, as AI scholars Kate Crawford and Meredith Whittaker have described, is largely internal to the process of data collection: errors in data collection, like inaccurate methodologies, can cause inaccurate depictions of reality.[14] This absence of representation is a profound cause of the risk of bias in AI. A second issue of bias comes from an external source. It happens when the underlying subject matter draws on information that reflects or internalizes some forms of structural discrimination and thus biases the data as a result.[15] Imagine, for example, a situation in which data on job promotions might be used to predict career success, but the data were gathered from an industry that systematically promoted men instead of women.[16] While the first kind of bias can often be mitigated by "cleaning the data" or improving the methodology, the latter might require interventions that raise complex political ramifications because of the structural nature of the remedy that is required.[17]

As a result, bias can surface in the context of input bias (when the source data are biased because they may lack certain types of information), training bias (when bias appears in the categorization of the baseline data), or through programming bias (when bias results from an AI system learning and modifying itself from incorporating new data).[18] In addition, algorithms themselves can also be biased: the choices that are made by humans – what features should be used to construct a particular model, for example – can comprise sources of inaccuracy as well.[19] An additional source of error can come from the training of the algorithm itself, which requires programmers to decide how to weigh sources of potential error.[20]

All the prior harms may seem representational in nature, but they cause discriminatory effects. If the prior discussion focused on the risks of exclusion from statistical and historical underrepresentation in a data set, there is also the opposite risk of overrepresentation, which can lead to imprecise perceptions and troubling stereotypes. In these instances, due in part to overrepresentation in the data set, an algorithmic model might associate certain traits with another unrelated trait, triggering extra scrutiny. In such cases, it can be hard to prove discrimina-

tory intent in the analysis; just because an algorithm produces a disparate impact on a minority group, it does not always mean that the designer intended this result.[21]

Even aside from concerns about data quality and representation, a second cluster of issues emerges from the intersection of privatization and AI-driven governance. Constitutional law scholar Gillian Metzger has presciently observed that "privatization is now virtually a national obsession."[22] Her work describes a foundational risk that private industry is taking the lead in designing modes of governance.[23] Notably, private contractors exercise a broad level of authority over their program participants, even when government officials continue to make determinations of basic eligibility and other major decisions.[24] These trends toward privatization and delegation are endemic throughout government infrastructure, and many draw on machine learning techniques.[25] As intellectual property law scholar Robert Brauneis and information policy law scholar Ellen Goodman have eloquently noted, "the risk is that the opacity of the algorithm enables corporate capture of public power."[26]

Today, algorithms are pervasive throughout public law, employed in predictive policing analysis, family court delinquency proceedings, tax audits, parole decisions, DNA and forensic science techniques, and matters involving Medicaid, other government benefits, child support, airline travel, voter registration, and educator evaluations.[27] The Social Security Administration uses algorithms to aid its agents in evaluating benefits claims; the Internal Revenue Service uses them to select taxpayers for audit; the Food and Drug Administration uses algorithms to study patterns of foodborne illness; the Securities and Exchange Commission uses them to detect trading misconduct; local police departments employ their insights to predict the emergence of crime hotspots; courts use them to sentence defendants; and parole boards use them to decide who is least likely to reoffend.[28]

As legal scholar Aziz Huq has explained, the state uses AI techniques for targeting purposes (that is, decisions on who to investigate or how to allocate resources like aid) and for adjudicatory purposes (in which the state may rely on AI techniques as a stand-in for a judicial determination).[29] To these two parameters, we might add on a third, involving AI-driven forensic techniques to aid the state in determining whether a legal violation has taken place: for example, machine learning techniques that analyze breath alcohol levels. In such cases, while AI might aid the state in gathering evidence, the ultimate determination of compliance (or lack thereof) may rest with human judgment. Here, the selection of a perpetrator might be performed by human law enforcement (who also determine whether evidence supports that a violation has taken place), but the evidence might be informed by an AI-driven technique.

Many of these tools are privately developed and proprietary. Yet the rise of proprietary AI raises a cluster of issues surrounding the risk of discrimination:

one involving the deployment of AI techniques by private entities that raises legal concerns; and another involving the deployment of AI techniques by public entities that raises constitutional concerns. Taken together, these systems can often impose disparate impacts on minority communities, stemming from both private and public reliance on AI. In one example from Pennsylvania, an automated system called the Allegheny Family Screening Tool was used to determine which families were in need of child welfare assistance. But the system entailed the risk of racial disparity: since Black families were more likely to face a disproportionately higher level of referrals based on seemingly innocuous events (like missing a doctor appointment), they were likely to be overrepresented in the data. Parents also reported feeling dehumanized within the system by having their family history reduced to a numerical score. Moreover, given the large amount of data the system processed (and the sensitivity of the data), it carried a serious risk of data breaches.[30]

Each of these prior concerns, as Huq points out, maps onto concerns regarding equality, due process, and privacy, and yet, as he notes, each problem is only "weakly constrained by constitutional norms."[31] Not only would it be difficult to determine whether someone's rights were violated, but parties who were singled out would find it difficult to claim violations of equality, due process, or privacy, especially given the deference enjoyed by the decision-maker.[32] Further, the opacity of these systems raises the risk of (what I have called elsewhere) "information insulation," which involves an assertion of trade secret protection in similar cases.[33]

Each layer of AI-driven techniques raises profound questions about the rule of law. Here, privatization and automation become intimately linked, often at the cost of fundamental protections, like due process. The problem is not just that governmental decision-making has been delegated to private entities that design code; it is also the reverse situation, in which private entities have significant power that is not regulated by the government. While the effects of algorithms' predictions can be troubling in themselves, they become even more problematic when the government uses them to distribute resources or mete out punishment.[34] In one representative case, a twenty-seven-year-old woman with severe developmental disabilities in West Virginia had her Medicaid funds slashed from $130,000 to $72,000 when the vendor began using a proprietary algorithm, making it impossible for her to stay in her family home.[35] When she challenged the determination on grounds of due process, the court agreed with her position, observing that the vendor had failed to employ "ascertainable standards," because it provided "no information as to what factors are incorporated into the APS algorithm," nor provided an "individualized rationale" for its outcome.[36] The district court concluded that the lack of transparency created an "unacceptable risk of arbitrary and 'erroneous deprivation[s]' of due process."[37]

As the previous example suggests, while automation lowers the cost of decision-making, it also raises significant due process concerns, involving a lack of notice and the opportunity to challenge the decision.[38] Even if the decisions could be challenged, the opacity of AI makes it nearly impossible to discern all of the variables that produced the decision. Yet our existing statutory and constitutional schemes are poorly crafted to address issues of private, algorithmic discrimination. Descriptively, AI carries similar risks of majoritarian control and systemic prejudice, enabling majority control at the risk of harming a minority. And yet our existing frameworks for regulating privacy and due process cannot account for the sheer complexity and numerosity of cases of algorithmic discrimination. In part because of these reasons, private companies are often able to evade statutory and constitutional obligations that the government is required to follow. Thus, because of the dominance of private industry, and the concomitant paucity of information privacy and due process protections, individuals can be governed by biased decisions and never realize it, or they may be foreclosed from discovering bias altogether due to the lack of transparency.

If we consider how these biases might surface in AI-driven decision-making, we can see more clearly how the issue of potential bias in AI resembles the very problem of majority control that Ely wrote extensively about, even though it involves privatized, closed, automated decision-making. If our systems of AI are driven by developers or trained on unrepresentative data, it feeds into the very risk of majoritarian control that judicial review is ideally designed to prevent. I want to propose, however, another story, one that offers us a different set of possibilities regarding the building of trust by looking, again, to the prospect of judicial review.[39] Here, I want to suggest that AI governance needs its own theory of representation-reinforcement, extending to every person within its jurisdiction the equal protection of the law, in essentially the same way that the Constitution purports to.

Where metrics reflect an inequality of opportunity, we might consider employing a similar form of external judicial review to recommend against adoption or refinement of these metrics. In doing so, an additional layer of judicial or quasi-judicial review can serve as a bulwark against inequality, balancing both substantive and process-oriented values. Here, we might use judicial review, not as a tool to honor the status quo, but as a tool to demand a deeper, more substantive equality by requiring the employment of metrics to address preexisting structural inequalities. And if filing an actual legal case in the courts proves too difficult due to an existing dearth of regulation, then I would propose the institution of independent, quasi-judicial bodies to ensure oversight for similar purposes.

What would a representation-reinforcement theory – or relatedly, a theory of judicial review – accomplish in the context of AI? While a detailed account of

representation and reinforcement is hard to accomplish in a short essay, I want to focus on two main sets of possibilities, the first stemming from Ely's concept of virtual representation. As I suggested earlier, one core issue with algorithmic decision-making is that it reflects an inherently regressive presumption: decisions, and data collected by past practices, adequately reflect – and predict – what we should do in the future, thereby "freezing" the possibility of a deeper and more meaningful form of substantive equality.[40] Unrepresentative data, in other words, can perpetuate inequalities through machine learning, leading to a feedback loop that further amplifies existing forms of bias.

Interestingly, Justice Stone and John Hart Ely identified roughly the same concerns regarding the lack of minority representation in the democratic pool, justifying a more aggressive form of intervention and oversight. In other words, just as Ely's theory predicts, disparities in representation – over- or underrepresentation – can fuel disparate results. Yet Ely's raising of the "judicial enforceable duty of virtual representation" enables us to see how profitably it can be recast to enfranchise the interests of minority populations in an AI-driven context. As Ely observed, one basic concern is that minorities must always be represented in the political process, and that we rely essentially on our judicial system to make sure that this happens.[41]

Here, one core element to accomplish this goal involves the necessity of creating a layer of institutional separation between the initial decision-maker (the AI system) and the reviewer (essentially, the system of judicial review). Like the division between the judiciary and the legislative branches, AI-driven systems can and must include systems of independent oversight that are distinct from the AI systems themselves. And there is evidence that this architectural solution is taking place. Consider an analogy from Europe's General Data Protection Regulation (GDPR), which requires separate data protection impact assessments (DPIA) whenever data processing "is likely to result in a high risk to the rights and freedoms of natural persons."[42] Large-scale data processing, automated decision-making, processing of data concerning vulnerable subjects, or processing that might prevent individuals from exercising a right or using a service or contract would trigger a DPIA requirement.[43] Notably, this model extends to both public and private organizations.[44]

One could easily imagine how this concept of independent review could be incorporated more widely into AI-driven systems to ascertain whether a system risks disparate impacts. A close look at these statements reveals a markedly thorough implementation of the concept of institutional separation: a DPIA statement is meant to be drafted by the organization's controller in order to show compliance with the GDPR; but the controller represents a separate entity from the organization processing the data.[45] In doing so, the system ensures a form of built-in virtual representation and review by putting the controller in the same position as

a judge to ensure compliance. Additional elements require an assessment of risks to individuals and a showing of the additional measures taken to mitigate those risks.[46]

Lastly, at present, as Ely suggests, judicial review is often necessary to ensure due process. Due process is especially needed in the context of AI so that individuals are able to ascertain the rationale behind AI-driven decisions and to guard against unclear explanations. In one case, in Houston, a group of teachers successfully challenged a proprietary algorithm developed by a private company, SAS, called the Educational Value-Added Assessment System (EVAAS) to assess public school teacher performance, resulting in the dismissal of twelve teachers with little explanation or context.[47] Experts who had access to the source code concluded that the teachers were unable to "meaningfully verify" their scores under EVAAS.[48] Ultimately, the court ruled against adopting use of the software because of due process concerns, noting, tellingly: "When a public agency adopts a policy of making high stakes employment decisions based on secret algorithms incompatible with minimum due process, the proper remedy is to overturn the policy."[49] Plainly, the court agreed with the due process concerns, noting that the generalized explanation was insufficient for an individual to meaningfully challenge the system's determination, and the case settled a few months later.[50]

The Houston case is instructive in underscoring the importance of safeguarding procedural protections like due process. Had it not been for the teachers' ability to bring this to a judicial forum to demand due process protection, the AI-driven injustice they faced would have never seen the light of day. By requiring AI systems to integrate similar entitlements of due process and independent oversight, we can ensure better outcomes and build more trust into the accountability of AI-driven systems overall.

In his essay forty years ago, Bolter predicted, "I think artificial intelligence will grow in importance as a way of looking at the human mind, regardless of the success of the programs themselves in imitating various aspects of human thought. . . . Eventually, however, the computer metaphor, like the computer itself, will simply be absorbed into our culture, and the artificial intelligence project will lose its messianic quality."[51]

We are still at a crossroads in adapting to AI's messianic potential. Ely wrote his masterful work at a time in which AI was just at the horizon of possibility. Yet the way that AI promises to govern our everyday lives mirrors the very same concerns that he was writing about regarding democracy and distrust. But the debates over AI provide us with the opportunity to elucidate how to employ AI to build a better, fairer, more transparent, and more accountable society. Rather than AI serving as an obstacle to those goals, a robust employment of the concept of judicial review can make them even more attainable.

AUTHOR'S NOTE

The author thanks Erwin Chemerinsky, James Manyika, and Neal Katyal for their insightful comments and suggestions.

ABOUT THE AUTHOR

Sonia K. Katyal is Associate Dean of Faculty Development and Research, Codirector of the Berkeley Center for Law & Technology, and Distinguished Haas Chair at the University of California, Berkeley, School of Law.

ENDNOTES

[1] J. David Bolter, "Artificial Intelligence," *Dædalus* 113 (3) (Summer 1984): 3.

[2] Ibid., 17.

[3] Ibid.

[4] *U.S. vs. Carolene Products Co.*, 304 U.S. 144, 153 n.4 (1938).

[5] Henry Paul Monaghan, "John Ely: The Harvard Years," *Harvard Law Review* 117 (6) (2004): 1749.

[6] Jane S. Schacter, "Ely and the Idea of Democracy," *Stanford Law Review* 57 (3) (2004): 740.

[7] John Hart Ely, *Democracy and Distrust: A Theory of Judicial Review* (Cambridge, Mass.: Harvard University Press, 1980), 103.

[8] Schacter, "Ely and the Idea of Democracy," 740.

[9] See Ely, *Democracy and Distrust*, 77, quoted in Schacter, "Ely and the Idea of Democracy," 741.

[10] See, for example, Lawrence Tribe, "The Puzzling Persistence of Process Based Theories," *Yale Law Journal* 89 (1063) (1980).

[11] See Schacter, "Ely and the Idea of Democracy," 760.

[12] See Andrew D. Selbst and Solon Barocas, "Big Data's Disparate Impact," *California Law Review* 104 (3) (2016): 688.

[13] Kate Crawford and Meredith Whittaker, "The AI Now Report: The Social and Economic Implications of Artificial Intelligence Technologies in the Near-Term" (New York: AI Now Institute, July 7, 2016, last modified September 22, 2016), 6–7, https://ainowinstitute.org/AI_Now_2016_Report.pdf [http://perma.cc/6FYB-H6PK (captured August 13, 2018)].

[14] Ibid.

[15] Ibid.

[16] Ibid., 6.

[17] Ibid.

[18] Nizan Geslevich Packin and Yafit Lev-Aretz, "Learning Algorithms and Discrimination," in *Research Handbook on the Law of Artificial Intelligence*, ed. Ugo Pagallo and Woodrow Barfield (Cheltenham, United Kingdom: Edward Elgar Publishing, 2018), 88–133.

[19] Michael L. Rich, "Machine Learning, Automated Suspicion Algorithms, and the Fourth Amendment," *University of Pennsylvania Law Review* 164 (4) (2016): 883–885.

[20] Ibid.

[21] See Joshua A. Kroll, Joanna Huey, Solon Barocas, et al., "Accountable Algorithms," *University of Pennsylvania Law Review* 165 (3) (2017): 693–694.

[22] Gillian E. Metzger, "Privatization as Delegation," *Columbia Law Review* 103 (6) (2003): 1369.

[23] Ibid., 1370.

[24] Ibid., 1387.

[25] David S. Levine, "The Impact of Trade Secrecy on Public Transparency," in *The Law and Theory of Trade Secrecy: A Handbook of Contemporary Research*, ed. Katherine J. Strandburg and Rochelle C. Drayfuss (Cheltenham, United Kingdom: Edward Elgar Publishing, 2012), 406–441.

[26] See Ellen P. Goodman and Robert Brauneis, "Algorithmic Transparency for the Smart City," *Yale Journal of Law and Technology* 20 (1) (2019): 109.

[27] Noting these areas of use, see "Litigating Algorithms: Challenging Government Use of Algorithmic Decision Systems" (New York: AI Now Institute, 2018), 5, https://ainowinstitute.org/litigatingalgorithms.pdf [https://perma.cc/KZ52-PZAH (captured January 22, 2019)]. For details on government uses of automated decision-making, see Danielle Keats Citron, "Open Code Governance," *The University of Chicago Legal Forum* 2008 (1) (2008): 356–357.

[28] Ronald Bailey, "Welcoming Our New Algorithmic Overlords?" *Reason*, October 1, 2016, https://reason.com/archives/2016/10/01/welcoming-our-new-algorithmic [https://perma.cc/YV7L-RK8N (captured August 23, 2018)].

[29] Aziz Z. Huq, "Artificial Intelligence and the Rule of Law," *University of Chicago Public Law and Legal Theory, Research Paper Series* 764 (2021): 3.

[30] Aziz Z. Huq, "Constitutional Rights in the Machine Learning State," *Cornell Law Review* 105 (7) (2020): 1893.

[31] Ibid.

[32] Ibid., 1894.

[33] See Sonia Katyal, "The Paradox of Source Code Secrecy," *Cornell Law Review* 104 (5) (2019): 1240–1241, https://scholarship.law.cornell.edu/clr/vol104/iss5/2/, citing David S. Levine, "Secrecy and Unaccountability: Trade Secrets in Our Public Infrastructure," *Florida Law Review* 59 (135) (2007): 111, http://www.floridalawreview.com/2010/david-s-levine-secrecy-and-unaccountability-trade-secrets-in-our-public-infrastructure/ (discussing the public interest concerns at stake).

[34] Hannah Bloch-Wehba, "Access to Algorithms," *Fordham Law Review* 88 (4) (2020): 1277.

[35] Ibid., 1277–1278, citing "First Amended Complaint for Injunctive & Declaratory Relief," *Michael T. v. Bowling*, No. 2:15-CV-09655, 2016 WL 4870284 (S.D.W. Va. Sept. 13, 2016), ECF No. 14.

[36] Ibid., 1278.

[37] Ibid.

[38] Ibid., 1249.

[39] Sandra Wachter, Brent Mittelstadt, and Chris Russell, "Bias Preservation in Machine Learning: The Legality of Fairness Metrics Under EU Non-Discrimination Law," *West Virginia Law Review* 123 (3) (2021): 17, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3792772.

[40] Ibid., 31.

[41] See Brian Boynton, "'Democracy and Distrust' after Twenty Years: Ely's Process Theory and Constitutional Law," *Stanford Law Review* 53 (2) (2000): 406.

[42] Andrew D. Selbst, "Disparate Impact in Big Data Policing," *Georgia Law Review* 52 (1) (2017): 170–171. See also "Data Protection Impact Assessments," United Kingdom Information Commissioner's Office, https://perma.cc/Q2NL-9AYZ (captured October 13, 2018).

[43] Ibid. Requiring DPIAs if the entity uses "systematic and extensive profiling or automated decision-making to make significant decisions about people," processes data or criminal offense data on a large scale, systematically monitors a publicly accessible place, processes biometric or genetic data, combines or matches data from multiple sources, or processes personal data in a way that involves online or offline tracking of location or behavior, among other categories.

[44] See Dillon Reisman, Jason Schultz, Kate Crawford, and Meredith Whittaker, "Algorithmic Impact Assessments: A Practical Framework for Public Agency and Accountability" (New York: AI Now Institute, 2018), 7, https://perma.cc/JD9Z-5MZC (captured October 13, 2018).

[45] Sagara Gunathunga, "All You Need to Know About GDPR Controllers and Processors," Medium, September 12, 2017, https://medium.com/@sagarag/all-you-need-to-know-about-gdpr-controllers-and-processors-248200ef4126 [https://perma.cc/8X46-8Y5D (captured October 13, 2018)].

[46] Ibid.

[47] See *Houston Federation of Teachers v. Houston Independent School District*, 251 F. Supp. 3d at 1168 (S.D. Tex. 2017); and Bloch-Wehba, "Access to Algorithms," 1282.

[48] See *Houston Federation of Teachers v. Houston Independent School District*, 251 F. Supp. 3d at 1168.

[49] See ibid., 1179.

[50] See Bloch-Webha, "Access to Algorithms," 1282–1283.

[51] Bolter, "Artificial Intelligence," 18.

Artificially Intelligent Regulation

Mariano-Florentino Cuéllar & Aziz Z. Huq

This essay maps the potential, and risks, of artificially intelligent regulation: regulatory arrangements that use a complex computational algorithm or another artificial agent either to define a legal norm or to guide its implementation. The ubiquity of AI systems in modern organizations all but guarantees that regulators or the parties they regulate will make use of learning algorithms or novel techniques to analyze data in the process of defining, implementing, or complying with regulatory requirements. We offer an account of the possible benefits and harms of artificially intelligent regulation. Its mix of costs and rewards, we show, depend primarily on whether AI is deployed in ways aimed merely at shoring up existing hierarchies, or whether AI systems are embedded in and around legal frameworks carefully structured and evaluated to better our lives, environment, and future.

Unheralded and by inches, computational tools clustered under the label "artificial intelligence" are creeping into state and U.S. federal agencies' toolkits for elucidating, implementing, and enforcing the law.[1] The Transportation Security Agency is required by law to deploy full-body millimeter-wave scanners trained to identify specific persons whose body shape indicates the need for further screening.[2] Sixty-three other civilian agencies of the federal government use more than 150 predictive tools to find facts, craft binding rules, exercise enforcement-related discretion, and detect violations of federal law.[3] Local and state governments use similar tools to detect employment-benefit fraud, predict child abuse, and allocate police.[4] In local criminal courts, prosecutors obtain convictions by drawing on probabilistic DNA analysis software.[5] Local, state, and federal governments also leverage regulation to induce private parties to create and adopt new computational tools. The Department of Health and Human Services in 2016 created an algorithmic "security risk assessment tool" for health care providers needing to verify that their medical-records systems comport with federal data-security rules.[6] Large investment banks increasingly adopt algorithmic tools as a means of complying with antifraud or money-laundering laws. Without fanfare, or wide public deliberation, the era of artificially intelligent regulation is almost certainly at hand.

We aim to map the potential, and risks, inherent in that new era. By *artificially intelligent regulation*, we mean regulatory arrangements that use a complex compu-

© 2022 by Mariano-Florentino Cuéllar & Aziz Z. Huq
Published under a Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) license
https://doi.org/10.1162/DAED_a_01920

tational algorithm, or another artificial agent, either to define a legal norm or to guide its implementation. To see how AI might be integrated into the regulatory process at four distinct points – problem identification, empirical inquiry, rule formulation, and rule implementation – consider the following examples:

- A statute governing financial institutions' anti–money laundering responsibilities might define an explicitly Bayesian learning tool as part of an adequate anti–money laundering system. So long as a bank incorporates the tool, it would fall into a safe harbor against liability. The legally mandatory instrument, moreover, would dynamically update to account for new sorts of malfeasance at the regulatory authority's direction.
- Selecting particular people or families for the nation's refugee resettlement program, an agency might adopt as regulation a machine learning instrument to make acceptable decisions accounting for more vectors than can be easily calculated by a human decision-maker. The instrument will once again dynamically update to account for changing patterns of migration, geopolitical conditions, climatic change, and regional economic conditions. Human decision-makers might have to overcome a variety of challenges to take account of all of those relevant and complex streams of information quickly and accurately. An AI instrument could account for this information in a manner that contrasts with how human decision-makers would approach the problem without wholly breaking from the forms of human decision-making.
- A pollution emissions standard for manufacturing plants might be enacted as a reinforcement-learning algorithm. This instrument would define targets based on changing patterns of behavior and calculations of elasticity. It would hence respond dynamically to changing circumstances, including shifting strategies by emitting companies and their customers, quicker and cheaper than human modifications of regulation.
- An AI-infused mechanism might be adopted by law to allocate vaccines during a pandemic based on evolving data about a disease's spread, its symptoms, and the public's behavioral responses. Again, the regulation would take the form of a reinforcement-learning tool that changed based on evolving public-health circumstances.

These examples share common traits. Critically, in each one, *the law itself operates through a legally preordained computation process*. In the first, regulation defines compliance in terms of a (continually updated) algorithm. In the second, the distribution of state benefits is a product of an algorithm cast in law; distributions are not set in advance but emerge as a result of the algorithm's interaction with novel information. In the third, the algorithm-as-law defines a standard of conduct for private parties accounting for ways in which they, and others, respond in

real time. In the final example, the regulatory goal (defined as, say, maximum epidemic abatement) is formulated by hand in advance, but how that goal is realized is constantly recalibrated via computation of new data. In each of these use cases, a machine substantially displaces a different sort of human judgment.

The ubiquity of AI systems in modern organizations all but guarantees that regulators or the parties they oversee will make use of learning algorithms or novel techniques to analyze data in the process of defining, implementing, or complying with regulatory requirements. At one end of the continuum is the relatively incidental, isolated use of an AI system to assess whether data indicate that organizational enforcement priorities have changed over the course of a decade. At the other end of the spectrum is the statute that defines a financial institution's responsibility to guard against money laundering by formally defining, as a legal norm instantiated in a digital medium, a specific Bayesian updating function. Somewhere on the continuum one might draw a line to distinguish "artificially intelligent regulation" from more incidental use of manufactured intelligence merely to offer limited advice to legal decision-makers or evaluate the implementation of ordinary laws.

What to make of these arrangements is an intricate question that merits no simple answer. The public debate on regulatory AI is polarized between boosters and doomsayers. AI's diffusion across state instrumentalities hence provokes either shrill alarm or unblinking optimism. Minneapolis, San Francisco, and Oakland, for example, have all banned private facial recognition technologies that trawl public surveillance footage with AI tools. These jurisdictions enact the view that "AI is invariably designed to amplify the forms of power it has been deployed to optimize."[7] In contrast, Chicago and Detroit recently purchased real-time facial recognition systems to integrate into their citywide camera networks.

We diverge from scholars who offer either pure celebration or lament about AI's effect on law. Rather, we think that artificially intelligent regulation holds promise and peril. As digitally native law, it exploits potential gains from new predictive technologies, and these gains have attendant costs and serious risks. We readily acknowledge AI's risks to human agency and democratic politics.[8] We also think that the environmental impact of an industry already producing an estimated 3–3.6 percent of global greenhouse gas emission will also loom larger as usage increases.[9] But we reject the broad claim that AI, as part of responsible social regulation with careful contingency planning and institutional safeguards, cannot deepen democracy, improve human welfare, or empower marginalized groups.[10] Its mix of ensuing harms and rewards will instead depend on whether AI is deployed merely to shore up existing hierarchies, or whether its use aims to empower and better our lives, environment, and future.

We offer here an account of the possibilities of artificially intelligent regulation as a good and as a harm. We then offer thoughts on the "metaregulation"

of artificially intelligent regulation – that is, the larger regulatory frameworks in which agencies' decisions to adopt or reject AI tools might be nested – within a democratic framework. Neither wholesale resistance nor an unthinking embrace of AI governance is justified. The national state and its agencies will almost certainly deepen entanglements with new predictive technologies. The ensuing form of artificially intelligent regulation, though, is not graven in stone. Experimentation with AI can help us better understand and resolve challenges arising from society's often-conflicting expectations of the legal system for technical accuracy, democratic legitimacy, even-handed enforcement, and the nuanced consideration of situational factors. These various rule-of-law elements can be in tension with each other. AI systems can relax that tension, or perhaps exacerbate it in a specific case. But we see no alternative to the hard work of making sure that artificially intelligent regulation is designed to, and in fact does, advance the common good, and not deepen inequality or short-circuit democratic judgment.

Artificially intelligent regulation (AIR) is a legal norm that directly incorporates an algorithm capable of learning and adapting to new information, or the closely related activity of relying heavily on an algorithm to interpret or enforce a regulatory norm that may or may not itself directly incorporate an algorithm. The agency problem in regulation is familiar, but the AIR solution for it – and potentially achieving other goals – is novel.

We focus here on "regulation" in the sense of laws, rules, and guidance promulgated by an agency or department as part of an overarching legal framework for private activities like financial trading or health care. Regulation also includes the government's efforts to control its own workings, such as policing and immigration. We do not address here the role of AI in the common law.[11] Our topic is distinct from discussions of "personalized" common-law rules of contract and tort law developed by courts rather than regulators.[12]

Our topic has analogies to certain long-standing arrangements in regulatory law. Some regulations already incorporate external standards by reference, such as industry norms, or encompass nontextual information.[13] Although current administrative norms governing the *Federal Register* (the authoritative compendium of all regulations promulgated by agencies of the national government) may complicate the inclusion of a dynamic algorithm directly in a federal regulatory rule through incorporation by reference, both statutes and regulatory rules are sometimes drafted to allow agencies or the public to take account of changing knowledge or conditions.[14] AIR can also act as a supplement or substitute for bureaucratically lodged discretion. The law is itself capable of evolving as agencies learn. Just as case-by-case adjudication elaborates the common law, so artificially intelligent law also adapts. But the locus of adaptation of AI is likely to be a standard internal to a statute or regulation, not a body of case law accreting over time.

Even well before legal norms become automated or intelligent, regulators will have little choice but to take seriously the world's increasing dependence on AI. The Internet shook governance beginning in the late twentieth century. It forced public agencies to contend, willingly or not, with new ways of disseminating information, networking computers, and shaping public perceptions.[15] Regulators cannot unwind the widespread commercial adoption of AI techniques, such as backpropagation, neural nets, and large-language models, among contemporary firms. Algorithmic social media feeds, big-data trading platforms, and medical diagnostic tools powered by machine learning are, moreover, unlikely to be abandoned given consumer demand and the real welfare gains derived from them. Nor will regulated firms, including media platforms, banks, hospitals, and manufacturers, cease to innovate in respect to these tools – if nothing else because of unstinting foreign competition. The synergies between state and private enterprise in China, in particular, lend this commercial contest a geopolitical edge that cannot be wished away.[16] Military agencies will keep pioneering technologies – like the communication protocols developed for the ARPANET project in the late 1960s that preceded the Internet – that invariably leak into civilian application. The conclusion that AI will increasingly infuse both government and society, therefore, is not mere lazy technological determinism. It is a reasonable inference from readily observable trends.

Still, invention is not the same as innovation.[17] Not all digital tools catch on. The recent history of machine learning innovation has been uneven, punctuated by unexpected stops and starts. Whether new technologies are picked up, and how their benefits and costs are distributed, depends on social, economic, cultural, and even legal forces. However acute the pressures toward AI diffusion and adoption might be at this moment, nothing excuses regulators, jurists, and scholars from the difficult task of figuring out how those new tools are slotted into, and interact with, existing private or public institutions, as well as extant hierarchies coded by race, ethnicity, gender, or wealth. Nothing makes existing technological arrangements ineluctable. The monopolistic scale and network effects of dominant social media platforms, for example, was a contingent result of federal regulatory choices.[18] Antitrust law might still find a way to reverse Facebook's and Google's dominance. Locally, the Los Angeles Police Department's April 2020 decision to abandon Palantir's crime-prediction software suggests that not all technological adoptions travel a one-way street. Predictions that AI inevitably serves to discriminate and disempower can enlist powerful historical examples. Their forward-looking force rests on a questionable disregard of democratic agency.

It would be a mistake to say that artificially intelligent regulation will ever completely displace human judgment in some form at some stage of the regulatory process. Human discernment designs and creates the learning tool embedded in the law. The fact that the application of rules to specific cases does not hap-

pen through the exercise of human discretion does not – indeed, cannot – lead to a complete absence of such discretion.[19] There may not always be a human in (or on) the loop, but there is always a human exercising her judgment as to policy goals, what data are relevant to those goals, and how best to reconcile competing values: she may simply not be visible. With AIR, those judgments likely occur earlier in the design and implementation process. These judgments will tend not to be situated decisions, of the kind regulators now make, about how a norm applies to specific facts and particular persons. An instrument for matching refugees, for example, will not have information on particular flows of people, and almost certainly will be designed by engineers with little or no direct understanding of the refugee experience. As a result, the design stage of artificially intelligent regulations and the ensuing specification of predictive tools is a context in which biases (including invidious beliefs about race, gender, or other legally protected classifications), blind spots, and inaccurate generalizations filter into law. This human element of artificially intelligent regulation may well be occluded from the view of regulated parties.

Nevertheless, policy-makers and the public may have compelling reasons to move human judgment upstream and to filter it through a machine learning tool. At a very general level, AIR has the potential to make law and legal instruments more trustworthy – more amenable to accounting and discipline – and thereby to reduce the transaction costs of translating legal norms across different platforms and institutions.

The positive case for AIR comprises several elements. First, AIR can push agencies to define a societal goal more explicitly. Many AI instruments are organized around a "cost function" that examines each set of predictions of an outcome variable derived from historical data and defines a "cost" or penalty between predictions and the true (observed) outcome. The instrument is then trained to minimize that cost.[20] Writing a cost function requires a precise understanding of the social goals regulation seeks to advance. Because that judgment must be explicitly made, the cost function is an opportunity to air to the public both regulatory goals and the manner in which trade-offs are made.

A second benefit of AIR is flexibility over time. Agencies presently promulgate regulations and guidance as a means of implementing statutes periodically enacted by a legislature. Regulation often uses abstract or vague terms, or simply broadly sets a policy goal. Implementing that abstract statutory ambition – whether it is a safe workplace, a technologically feasible but environmentally tolerable level of emissions, or a decent refugee regime – requires translation. Regulators need then to write out their abstract goals in terms of particular rules or applications, bringing lofty aspiration into material form. AIR allows a well-informed legislature to install into law its abstract policy goal in a durable and adaptable way. Where regu-

lation adopts a reinforcement learning tool, the legislature also benefits from information that is not available at the time a law is passed. Hence, a resettlement algorithm might account for unanticipated shifts in migration patterns, or an antifraud tool could learn to recognize new species of criminal conduct. Thanks to this ability to build into law the capacity to dynamically update, a legislature condemned to only intermittent formal action via bicameralism and presentment is freed from frozen-at-the-moment-of-enactment text. This kind of flexibility may be especially valuable if the U.S. Supreme Court imposes new restrictions on Congress's ability to delegate through general grants of powers to federal agencies, with the latter filling in details with regulations.[21] A law directing creation of an AIR might be a substitute for flexibility otherwise exercised through agency rulemaking over time.

A related benefit pertains to legislators' "agency cost" problem. Regulators may have different policy preferences from legislators. They might be excessively close to a regulated industry. Or they might slack off.[22] One way to mitigate agency slack is with *ex post* judicial review. But the use of courts as an oversight mechanism has costs. Litigation can be used to delay desirable regulation. Fearing a suit, budget-constrained agencies might forgo action. Regulated parties, anticipating judicial review, have an incentive to lobby for particular judicial appointments.[23]

AIR addresses agency slack in a different way. By impounding their judgments into a digitally native tool, legislators drain away later discretion about how a law is enforced. The resources used up in translating verbal standards back and forth to code and mathematical specification are likely to be smaller than the social resources sucked up by litigation clashes between interest groups and the government. AIR, however, does not eliminate agency problems entirely. Realistically, legislators must rely on technologists and coders to craft an instrument. Unless a legislator can trust the designers of digitally native law, as well as the sources of training data, the specter of "capture" and distorted preferences arises once more.[24] Legislators could demand benchmarking and transparency in AI design "appropriate for practical reasoning," not just in terms of technical detail.[25] Such arrangements might further facilitate either *ex post* judicial review (especially when individual rights are at issue) or legislative committee-based oversight.

The advance of AIR under these conditions may also alter democratic governance more broadly. At present, a legislature enacts a law with limited control over how its terms are understood and applied in the future. Later legislatures can use their appropriations power and their ability to jawbone agency leaders to nudge regulation toward their preferred policies, even when they diverge from those of the enacting legislature. Sometimes, courts step in to interpret statutes in ways that force the later legislature to act more overtly by passing new law.[26] But not always, and not reliably.

AIR might scramble such arrangements. In principle, it empowers an enacting legislature. That body has the ability to enact not just the law in an abstract form,

but also to embed a mechanism for updating. This sharpens the importance of the discrete political moment in which a law is enacted; it also diminishes the importance of the legislative power to influence agencies in the long term. Arguably, this is salutary in terms of democratic norms. It ups the stakes of the actual legislating moment, when the voting public is most likely attentive, while diminishing the importance of periods in which the public is less engaged, and legislative influence more diffuse. This helps voters exercise retrospective judgment about their representatives. On the other hand, AIR, in a paradoxical sense, by making formal laws less brittle and more capable of built-in adaptation, could conceivably enable long-past legislative coalitions to endure beyond their expiration date. Hence, it may empower the dead hand of the past against the influence of living legislators wielding a current democratic imprimatur.

Finally, it is worth considering whether AIR can be used to broaden access to legal institutions and the benefits of law more generally. Algorithmic tools already facilitate estate planning via websites such as LegalZoom. While these instruments are not without complications, it is worth considering ways in which AIR might be used to empower ordinary citizens presently discouraged from seeking legal remedies by litigation's complexity and cost.[27] This is one important way of resisting the complacent assumption that AI is an innovation that necessarily and inevitably concentrates power and increases pernicious social inequalities.

All should not be presumed to be well with this potential new era of regulation. Just as it enables optimal adaptability, diminished agency costs, and lower transaction costs, so artificially intelligent regulation will engender new problems of transparency, legitimacy, and even equity. All raise fundamental questions of constitutional magnitude.

To begin, it is premature to assume AIR always reproduces undesirable or malign forms of hierarchy. Though regulation is not guaranteed to enhance social welfare, neither is it intrinsically regressive. It has advanced the cause of civil rights, workplace safety and health, environmental protection, and consumer rights. AIR is just one species of regulation. Of course, all lawmaking risks interest-group capture or the unintended perpetuation of invidious stereotypes. AIR, like any kind of legal intervention, must be scrutinized for those risks. In particular, AIR empowers a new class of experts – computer scientists and engineers – at present noteworthy for its lack of gender, racial, and ethnic diversity. Finding diversity in such expertise and turning the latter to serve the public good is not impossible: biological and medical science has shown as much. But it will require sustained institutional change.

More seriously, the ends and means of AIR – like many of the complex statutes that Congress, in particular, has enacted – are not necessarily readily perceived or understood by nonspecialist members of the public or elected officials. The value

of an algorithm that diminishes conventional principal-agent problems involving human-led agencies also means that standards might evolve in problematic ways. This may be a result, for example, of mistakes in how a reinforcement learning reward function is specified, or it can be a consequence of adversarial disruptions. There is a question whether any adaptational "drift" distorts what the law achieves, or instead demands fresh involvement by the very mix of experts, politically accountable officials, and competing stakeholder pressures that optimistic proponents might expect these new forms of law to render unnecessary.

A yet more fundamental question is whether an AI-based legal arrangement would be perceived as legitimate in either a sociological or legalistic sense. The ability of the public to understand what AI does at the front end is limited, although that is also true of many existing laws and legal institutions. Leaving aside the precision with which a dynamic legal provision "aligns" with a defensible macroconcept of social welfare in advance, the way such provision *evolves* over time is not made legitimate without further ingredients. These include the capacity of concentrically larger circles of people, including agency officials and regulated parties at minimum, to understand certain things about how a system performs. Also relevant is affected parties' capacity to argue in terms the public understands about why AIR is performing adequately (or not) relative to the rest of the jurisdiction's legal commitments. Agencies or lawmakers could also create "tripwires" to prevent excesses in the use of public, coercive authority; capture or co-option by private interests through de facto private delegation; or violations of due process, equal protection, or anticorruption norms. Certain uses of coercion may also be *ipso facto* illegitimate without human oversight.

Such measures could be calibrated to promote institutions that allow debate about how a law gets implemented in a particular situation and about the policy and value assumptions supporting the law. Equally important are arrangements that prevent the use of AIR as a shield to prevent public accountability for the coercive use of power. Here, "public accountability" means that some people must accept responsibility for the use of coercive authority in ways that account for material and emotional consequences, including loss of income, reputational degradation, loss of interesting work, and misrecognition by peers or authority figures.

Finally, with opacity comes the risk that algorithms reenact malign hierarchies of race, ethnicity, class, and gender via inscrutable code and invisible design choices.[28] The terminology of "bias" in AI is used in varied and inconsistent ways. In our view, the most powerful normative concerns arise when the use of AI imposes material harms on a historically subordinated group.[29] Machine bias defined in this way need not flow from any conscious decision to suppress a historically subordinate group. It can result simply from inattention or ignorance by programmers who are not members of those classes. Preventing intentional or inadvertent reproduction of these hierarchies requires active attention to the code

inserted into regulation. As recent turmoil at Google's ethics division suggests, the implementation of equity is no simple matter, but demands organizational leadership and effective staffing.

None of this means AIR should be eschewed. But technical limitations and public resistance mean AIR will likely be limited in scope for some time. More interesting to us is how the emergence of AIR raises questions about the "metaregulatory" structure of administrative and regulatory law. That is, how should the law itself guide the creation and oversight of digitally native law?

The law has already developed tools to audit and evaluate ordinary regulation; cost-benefit analysis is foremost among them. AIR requires rethinking and retooling government's auditing and oversight capacity to extend values of equity and rationality into its frameworks. As governments create "sandboxes" in which to build and test AIR, they will need to apply robust norms of transparency and benchmarking to ensure that AI is not just the product of – but also facilitates – reasonable and informed deliberation. Experiments with AIR may benefit from building in some of the encumbrances that surely make laypeople wonder about law as it operates today. Digitally infused regulations might therefore explicitly incorporate interpretive mechanisms that will "translate" a standard into ordinary language. Periodic audits for practical bias along race, gender, and other lines might be mandated by law, with failure to pass connected to a penalty of statutory rescission.

More broadly, a federal agency can be imagined for being responsible for sourcing, testing, and auditing new digital tools. Such an agency could be especially helpful given some of the persistent difficulties public organizations face when making procurement decisions. The agency would benefit from a capacity to experiment with recruitment and retention tools, including rotation and part-time arrangements subject to appropriate safeguards against conflicts of interest, to layer into the highest levels of the public sector the kind of expertise and mix of cultures helpful in enhancing government capacity for assessment of AIR. It would operate much as the General Services Administration, established by President Harry Truman in 1949, serving as a hub for digitally native law, a source of auditing expertise, and a locus for public complaint.

For the foreseeable future, AIR offers fascinating possibilities for enhancing governance, but it will nonetheless face intense constraints. Given the risks entailed, perhaps this is as it should be. If AIR is to become legitimate, it must face a trial by fire under the abiding rule-of-law constraints familiar from our existing, imperfect legal system. Further, it will be subject to the coterie of pluralistic pressures capable of creating such enormous friction for even the most elegantly designed legal reforms. Both will confer legitimacy and limit risks of severe

error, but also erode AIR's possibilities and promise. Perhaps such friction is not entirely useless. Perhaps, indeed, it has the potential to force nuance into discussions about how to reconcile contending ideas about what sort of social welfare regulation is supposed to advance. The resulting constraints also offer a powerful reminder that the social benefits of AIR depend at least as much on our society's capacity to engage in intelligent governance as they do on continued progress in machine learning.

ABOUT THE AUTHORS

Mariano-Florentino Cuéllar, a Fellow of the American Academy since 2018, is President of the Carnegie Endowment for International Peace. A former Justice of the Supreme Court of California, he served in the Obama administration as Special Assistant to the President for Justice and Regulatory Policy at the White House Domestic Policy Council. He was previously the Stanley Morrison Professor of Law at Stanford University and is the author of *Governing Security: The Hidden Origins of American Security Agencies* (2013).

Aziz Z. Huq is the Frank and Bernice J. Greenberg Professor of Law at the University of Chicago Law School. He is the author of *The Collapse of Constitutional Remedies* (2021), *How to Save a Constitutional Democracy* (with Tom Ginsburg, 2018), and *Unbalanced: Presidential Power in a Time of Terror* (with Frederick A. Schwarz, 2007).

ENDNOTES

[1] David Freeman Engstrom, Daniel E. Ho, Catherine M. Sharkey, and Mariano-Florentino Cuéllar, *Government by Algorithm: Artificial Intelligence in Federal Administrative Agencies* (Washington, D.C.: Administrative Conference of the United States, 2020).

[2] *Federal Register* 81 (42) (2016): 11363.

[3] Engstrom et al., *Government by Algorithm*, 15–16.

[4] Virginia Eubanks, *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor* (New York: Macmillan, 2018).

[5] *People v. Chubbs*, 2015 WL 139069 (January 9, 2015).

[6] "Security Risk Assessment Tool," Health IT, https://www.healthit.gov/topic/privacy-security-and-hipaa/security-risk-assessment-tool.

[7] Kate Crawford, *Atlas of AI* (New Haven, Conn.: Yale University Press, 2021), 224.

[8] Shoshana Zuboff, *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power* (New York: Public Affairs, 2019).

[9] Lotfi Belkhir and Ahmed Elmeligi, "Assessing ICT Global Emissions Footprint: Trends to 2040 & Recommendations," *Journal of Cleaner Production* 177 (1) (2018): 448–463.

[10] Crawford, *Atlas of AI*, 223.

[11] Mariano-Florentino Cuéllar, "A Common Law for the Age of Artificial Intelligence," *Columbia Law Review* 119 (7) (2019): 1773–1792.

[12] Ariel Porat and Lior Jacob Strahilevitz, "Personalizing Default Rules and Disclosure with Big Data," *Michigan Law Review* 112 (8) (2014): 1417–1478.

[13] Federal regulations may incorporate "published data, criteria, standards, specifications, techniques, illustrations, or similar material." 1 CFR § 51.7 (a)(2)(i); and Emily S. Bremer, "On the Cost of Private Standards in Public Law," *Kansas Law Review* 63 (2015): 279, 296.

[14] Office of the Federal Register, *IBR Handbook* (Washington, D.C.: Office of the Federal Register, 2018). See, for example, *Whitman v. American Trucking Associations*, 531 U.S. 457 (2001), discussing the EPA's responsibility under Section 109(b)(1) of the Clean Air Act to set National Ambient Air Quality Standards that are "requisite to protect the public health."

[15] See Jonathan Zittrain, *The Future of the Internet and How to Stop It* (New Haven, Conn.: Yale University Press, 2008).

[16] Aziz Z. Huq and Mariano-Florentino Cuéllar, "Privacy's Political Economy and the State of Machine Learning," *NYU Annual Survey of American Law* (forthcoming).

[17] David Edgerton, *The Shock of the Old: Technology and Global History Since 1900* (London: Profile Books, 2011).

[18] Dina Srinivasan, "The Antitrust Case against Facebook: A Monopolist's Journey towards Pervasive Surveillance in Spite of Consumers' Preference for Privacy," *Berkeley Business Law Journal* 16 (1) (2019): 39–99.

[19] Aziz Z. Huq, "A Right to a Human Decision," *Virginia Law Review* 106 (3) (2020): 611–688.

[20] David Lehr and Paul Ohm, "Playing with the Data: What Legal Scholars Should Learn about Machine Learning," *UC Davis Law Review* 51 (2) (2017): 653–717.

[21] Cass R. Sunstein, "The American Nondelegation Doctrine," *George Washington Law Review* 86 (5) (2018): 1181–1208.

[22] Kenneth J. Meier and George A. Krause, "The Scientific Study of Bureaucracy: An Overview," in *Politics, Policy, and Organizations: Frontiers in the Scientific Study of Bureaucracy*, ed. George A. Krause and Kenneth J. Meier (Ann Arbor: University of Michigan Press, 2003): 1–19.

[23] Cass R. Sunstein, "On the Costs and Benefits of Aggressive Judicial Review of Agency Action," *Duke Law Journal* 1989 (3) (1989): 522–537.

[24] Justin Rex, "Anatomy of Agency Capture: An Organizational Typology for Diagnosing and Remedying Capture," *Regulation & Governance* 14 (2) (2020): 271–294.

[25] John Zerilli, John Danaher, James Maclaurin, et al., *A Citizen's Guide to Artificial Intelligence* (Cambridge, Mass.: MIT Press, 2021), 33.

[26] Einer Elhauge, "Preference-Eliciting Statutory Default Rules," *Columbia Law Review* 102 (8) (2002): 2162–2290.

[27] Emily S. Taylor-Poppe, "The Future Is Complicated: AI, Apps & Access to Justice," *Oklahoma Law Review* 72 (1) (2019): 185–212.

[28] Safiya Umoja Noble, *Algorithms of Oppression: How Search Engines Reinforce Racism* (New York: NYU Press, 2018).

[29] Aziz Z. Huq, "Racial Equity in Algorithmic Criminal Justice," *Duke Law Journal* 68 (6) (2019): 1093–1134.

Socializing Data

Diane Coyle

Will the proliferation of data enable AI to deliver progress? An ever-growing swath of life is available as digitally captured and stored data records. Effective government, business management, and even personal life are increasingly suggested to be a matter of using AI to interpret and act on the data. This optimism should be tempered with caution. Data cannot capture much of the richness of life, and while AI has great potential for beneficial uses, its delivery of progress in any human sense will depend on not using all the data that can be collected. Moreover, the more digital technology rewires society, creating opportunities for the use of big data and AI, the greater the need for trust and human deliberation.

Data have always been important for government and policy. Statistics are, as the name suggests, categorized data useful for states.[1] States have collected and collated data for centuries, not least for the purposes of taxation. Censuses too are ancient, defining the boundaries of power, though they are likely to be replaced by other government-collected data sets about individuals.[2] The purpose of governmental measurement is to create conceptual order, to classify the vast array of possible data points into meaningful categories, enabling better decisions. Over the quarter-millennium of modern economic growth, the scope of data collection and processing into statistics has become increasingly extensive.

In *Seeing like a State* (1998), political scientist James Scott argues that modern states classify reality to improve the *legibility* of what they govern, to better control it. He writes: "Legibility implies a viewer whose place is central and whose vision is synoptic. . . . This privileged vantage point is typical of all institutional settings where command and control of complex human activities is paramount."[3] Many of his examples of states bending reality into order concern economic activities such as forestry or agriculture, with reality conforming increasingly to the classifications devised to understand it. There is a feedback loop whereby statistics collect and classify data points found in the wild, then subsequently influence activities and shape reality over time, so that future data will be more likely to fit into the predefined categories. This has been described by statistician André Vanoli as "the dialectic of appearance and reality."[4] Or as historian Theodore Porter put it, "The quantitative technologies used to investigate social and economic life always work best if the world they aim to describe can be remade in their image."[5]

© 2022 by Diane Coyle
Published under a Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) license
https://doi.org/10.1162/DAED_a_01921

For example, the principal measure of economic progress since the early 1940s has been gross domestic product (GDP).[6] Governments gear their policies toward increasing GDP, and people duly respond to the incentives created by policies such as tax breaks, subsidies, public infrastructure investment, or cheaper meals out.[7] Disappointing statistics can topple governments, as they did with the UK Labour government of the late 1970s, paving the way for the Thatcherite revolution. GDP has not been a terrible metric for progress: compared with previous generations, our living standards are without doubt higher. We have better health, more leisure, more comfortable homes, and the convenience of many new technologies. Yet even at the dawn of GDP's invention, some realities had to be bent to fit the statistical framework. Some were rendered invisible, defined as being outside "the economy," such as household work and nature. Without nature, there is no economy and yet the consequences for sustainability of this fateful definitional choice are becoming all too clear, and the progress we thought we had is at least partly illusory.

Reality and the statistical picture also diverge when reality is changing. As statistician Alain Desrosières has written, "In its innovative phase, industry rebels against statistics because, by definition, innovation distinguishes, differentiates and combines resources in an unexpected way. Faced with these 'anomalies,' the statistician does not know what to do."[8] At present, for official statisticians, life is one damned anomaly after another. For just as agriculture's share was overtaken by manufacturing in the industrial revolution, the material economy is smaller now relative to the dematerializing economy of digitally enabled services.[9] The statistical categories no longer fit well. Paradoxically, in the economy of ever more data, it is proving increasingly difficult to bring informational order, for the state to gain that desired legibility.

This is a paradox because the promise of big data and its use in AI has inspired renewed visions inside government of enhanced legibility. Such visions are not new. From the late 1950s onward, computers have seemed to promise a clearer, synoptic understanding of society.[10] One ambitious 1970s project was Project Cybersyn in Salvador Allende's Chile, administered by cyberneticist Stafford Beer, which was intended to implement an efficiently planned economy.[11] A similar vision of data-enabled, improved legibility has revived in the big data digital era. On the left of UK politics it found expression as "fully automated luxury communism."[12] In the UK Conservative government elected in 2019, it took physical shape as a control room at the heart of government, and a UK Strategic Command contract with tech firm Improbable to build a "digital twin," a simulation of the whole of Britain.[13] The fact that both ends of the political spectrum envision data-driven efficiency suggests a big data rerun of the 1930s socialist calculation debate.[14]

The thing that is seen in seeing rooms of these kinds – physical rooms with displays of information to inform decision-makers – is ordered data. There is a kind of commodity fetishism regarding the mechanics of displaying the data. The technology of data has long been glamorous, arousing intense public and political interest. The great exhibitions and world's fairs of the nineteenth and early twentieth centuries had popular displays of high-tech data management artifacts such as filing cabinets and cards.[15] The same is true of digital technology and Silicon Valley, which have inspired numerous nonfictional and fictional accounts. Databases have changed form over time as physical hardware and computational power have evolved, so the embodiment and the usability (searchability) of data have not been constant, and the technologies of display combined with the classification and conceptual framework organizing the data affect the way decision-makers understand the world. The emphasis on the synoptic view – through a computer simulation, through a room kitted out with the latest screens and data feeds – is an assertion of political control through greater legibility. Then–UK government adviser Dominic Cummings presented it as a matter of public interest:

> There is very powerful feedback between: a) creating dynamic tools to see complex systems deeper (to see inside, see across time, and see across possibilities), thus making it easier to work with reliable knowledge and interactive quantitative models, semi-automating error-correction etc, and b) the potential for big improvements in the performance of political and government decision-making.[16]

In other words, the claim is that data science and AI, suitably embodied in a seeing room, can be the vehicle for delivering "high performance" by government.

However, the emphasis is on the technologies of cognition and management, rather than the construction of the data going into the process, or the assessment of what constitutes improvement. The implicit assumption is that this is a determination made by the center, by those in the seeing room. This assumption is exactly why an ambition to use data for progress can embed biases, create ambiguity about accountabilities, or appear to be part of the surveillance society.[17] There is certainly nothing new about state attempts to exercise comprehensive surveillance. East Germany's Stasi offers an extreme recent example. Its data took analog form with a technological infrastructure turning data into seeing: card records with a bespoke filing cabinet technology, photographs, steam irons for opening mail, tape recordings, and computers. Despite the existence of formal regulations controlling access to this data, a citizen of the former German Democratic Republic was a *gläserne Mensch*, a transparent being. Perhaps we are all becoming transparent now. Digital technology makes the amassing of data records trivially cheap and easy by comparison with the 1980s, and security agencies have been doing this at scale.

Big tech companies, not just security agencies, have been amassing the biggest and best databases and the know-how to use them for a purpose. Their purpose is

profit, rather than public good, and their market power ensures they do not need to serve the interests of their users or the public in general. Big Tech's success vis-à-vis state power is amply evident in the erosion of national tax bases as ever more economic activity goes online. It is not clear how much governments can limit this.[18] As being able to raise tax revenues is a core state function, there can be little question about the power of the biggest digital companies. If the synoptic view of what is happening anywhere is available to anybody, it is Google or Facebook. They, not officials or politicians, are collecting, categorizing, and using the new proliferation of data.

As long as data are seen as individual property amenable to normal market exchange, that will continue to be the case, despite recent regulatory moves in several jurisdictions to enforce some data sharing by the tech giants. The reason big tech companies have been able to acquire their power is the prevailing conceptual framework, crystallized into law, for understanding data as property. Rather than the appreciation that data reflect constructed categories, a particular lens or framework measuring and shaping reality, data are seen as the collection of natural objects: the classifications codified and programmed into data feeds just are what they are. These constructed data records are then subject to legal rules of ownership. Data are presumed to be transferred and owned by corporations as soon as the user of a service has accepted its terms and conditions.

The consequences of this property rights concept applied to data, or information, illuminate why it is so pernicious. For example, John Deere and General Motors (as corporate persons) have claimed in U.S. copyright courts that farmers or drivers who thought they were purchasing their vehicles do not in fact own them and have no right to repair them. The company's reasoning is that a tractor is no longer mainly a metal object whose ownership as a piece of property is transferred from John Deere to the farmer, but rather an intangible data-fed software service licensed from the company, which just happens to have a tractor attached.[19] Indeed, screens with data about weather, soil conditions, and seed flow proliferate inside tractor cabins and feed into the diagnostic software installed by the manufacturer, which provides information to enable decisions raising crop yields. The John Deere claim to ownership of the intangible dominates the farmer's claim to ownership of the physical vehicle it is bundled with. To date, the courts have been largely sympathetic to the corporations and to the strong ownership claims made by Amazon over e-books, by makers of games on consoles, as well as by vehicle manufacturers.

One response to such corporate ownership over data and data processing claims has been the demand for corporations to pay for "data as labor."[20] With this, each data point an online business collects from users' activities would be rewarded with a small financial payment. However, as economist Zoë Hitzig and colleagues point out, this remedy also considers data as a transferrable, individual item of property, and implicitly as a natural object "given" by the underlying reality.[21]

The data-as-property perspective assumes data are an object in the world, with an independent reality, differing from other givens only in being intangible. Yet not only are data nonrival (their use does not deplete them so many people can use them), but they are also inherently relational. Data are social. Even when it comes to data that are seemingly ultrapersonal – for example, that I passed a particular facial recognition camera at a given moment – the information content and usefulness of the data are always relational.[22] A facial image needs to be compared with a police database. Even then its utility for the purpose of detecting suspected criminals depends on the quality of the training data used to build the machine learning algorithm, including its biases, the product of a long history of unequal social relations. The relational character of data means they are both constructed by social relations and a collective resource for which market exchange will not be the best form of organization.[23] Indeed, this is why there are few markets for data; where data are sold – for example, by credit rating agencies – the market is generally thin, with no standardized, posted prices. The use value of data – their information content enabling decisions to be made – is highly heterogeneous.

That markets are a poor organizational model for the optimal societal use of data is Economics 101. Does that make government the right vehicle to use big data and AI for the public good? Can and should governments aim to beat big tech at the seeing game? The promise of automating policy through seeing rooms and use of AI is greater efficiency and, potentially, better outcomes. Yet there is increasing use of algorithmic processes in arenas in which decisions could have a large impact on people's lives, such as criminal justice or social security.

Much of the literature on the informational basis of organizations focuses on complexity as the constraint on effective information-processing, given an objective function.[24] Automation is superior in routine contexts: more reliable, more accurate, faster, and cheaper. What is more, machines deal more effectively with data complexity than humans do, given our cognitive limitations. This is a key advantage of machine learning systems as the data environment grows more complex. The system is better able than any human to discern patterns and statistical relationships in the data, and indeed the more complex the environment, the greater the AI advantage over human-scale methods. However, whenever there is uncertainty, the advantage tips back to humans. The more frequently the environment changes in unexpected ways, or the more dramatic the scale of change, the greater the benefits of applying human judgment. The statistical relationships on which automated decision rules are based will break down in such circumstances (in economics this is known as the Lucas critique).[25] The selection of a machine or human to make decisions is generally presented as a trade-off. However, it has long been argued, or hoped, that AI can improve the terms of this trade-off.[26]

There are several reasons to doubt this. One is the well-known issue of bias in training data sets, the inevitable product of unfair societies in the way data are classified, constructed, collected, and ordered.[27] Any existing data set reflects both the classification framework used and the way that framework has shaped the underlying reality over time (that is, André Vanoli's dialectic referred to earlier). The data science community has become alert to this challenge and many researchers are actively working on overcoming the inevitable problems raised by data bias. But bias is not the only issue.

Another less well-recognized issue (at least in the policy world) is that decisions based on machine learning need an explicitly coded objective function. Yet in many areas of human decision-making – particularly the most sensitive, such as justice or welfare – objectives are often left deliberately implicit. Politics in democracies requires compromise on high-level issues so that low-level actions can be taken. These "incompletely theorized agreements" are not amenable to being encoded in machine learning (ML) systems, in which precision about the reward function is needed even if conflicting objectives are combined with different weights.[28] The further deployment of ML in applied policy practice may require more explicit statements of objectives or trade-offs, which will be challenging in any domain where people's views diverge.[29] There could be very many of these, even in policy areas that seem straightforward. For example, how should public housing be allocated? There has been a pendulum swinging over time between allocation based on need and allocation based on likelihood to pay rent. These are conflicting objectives, and yet many of the same families would be housed under either criterion.

The extensive discussions of value alignment in the AI literature tussle with how to combine the brutally consequentialist nature of AI with ambiguity or conflicts about values. Given any objective or reward function, ML systems will game their targets far more effectively than any bureaucrat ever did. All the critiques of target setting in the public management literature, on the basis that officials game these for their personal objectives, apply with extra force to systems automating target delivery. This has led to concerns – albeit overstated – about runaway outcomes far from what the human operators of the system wanted.[30] One possible avenue is inverse reinforcement learning – that is, when ML systems try to infer what they should optimize for – which can accommodate uncertainty about the objective, but takes the existing environment as the desired state of affairs.[31] Political theorist and ethicist Iason Gabriel rightly emphasizes the need for legitimate societal processes to enable value alignment; but we do not have these yet.[32]

Market arrangements based on the concept of private property transactions are inappropriate for data, given their relational characteristics. In economic terms, there are large externalities, whereby one individ-

ual's provision of data can have either negative (loss of privacy) or positive (useful information) implications for other people.[33] Rather than being considered as property amenable to market exchange, data instead need to be subject to governance arrangements of permitted access and use. Online, the offline norms of sociologist Georg Simmel's concept of "privacy in public" do not exist.[34] This concept refers to the norms people adopt limiting what they know about each other in their different roles. Even publicly available information (such as where somebody lives) is not made known in a specific context (such as the marking of an exam paper by their lecturer). These voluntary informational restraints and social relations of trust play an important role in sustaining desirable outcomes such as fairness, privacy, or self-esteem.[35] Similar norms do not exist online. Big tech joins up too many data about each of us. People can reasonably be concerned about government seeing rooms doing the same.

At the same time, some joining up of data for some uses could without question lead to improved outcomes for individuals. So we have ended up in the worst of all worlds: a "surveillance state" or "surveillance economy" in which valid privacy concerns about certain data uses prevent other uses of "personal" data for collective and individual good. Consider the successful argument that governments should not use data from COVID-19 apps to trace individuals' contacts during the pandemic, leading almost all governments to adopt the Google and Apple application programming interfaces (API) with privacy enforced, all the while as personal liberty was infringed through lockdowns tougher than would have been needed with effective contact tracing. Meanwhile, governments and researchers have been able to use big data and machine learning to inform policies during the pandemic but could have done much more to avert unequal health outcomes with linked data about individuals' health status, location, employment, ethnicity, and housing.

The debate about privacy has become overly focused on individual consent and data protection. It should be a debate about social norms and what is acceptable in different contexts, translated into rights of access and use for limited, specific purposes.[36] In both the commercial and the public sphere, the promise of AI for decision-making will not be realized unless the kind of information norms that operate offline are created online. The control of access and use is not just a technical issue but a social and political one.

As the world gets both more complex and more uncertain, big data and AI will need to socialize in another way, by combining with human judgment more often. The experiences of 2020, or the impact of extreme climate-related events from California burning to Texas freezing, are suggestive of the prospect that "radical uncertainty" will characterize the twenty-first century.[37] Anybody with any knowledge of forecasting (no matter how small or big the

data set) will know that uncertainty about future outcomes multiplies over time. "Further computational power doesn't help you in this instance, because uncertainty dominates. Reducing model uncertainty requires exponentially greater computation."[38]

As radical uncertainty increases, the digital transformation is meanwhile expanding the domain of human judgment and trust. Institutional economics has generally considered two modes of organization: the market, in which price allocates resources, and the hierarchy, in which authority and contract apply. But neither price nor authority function well as allocation mechanisms when knowledge-based assets are important in production.[39] That the market is a poor vehicle for the use and provision of public goods such as knowledge is a standard piece of economic theory. Similarly, a large body of management literature notes that knowledge is hoarded at the top of hierarchical organizations, which are consequently good at routine activities but not at adaptation or innovation.

Trust is a more effective mechanism than either market exchange or command-and-control for coordinating knowledge-intensive activities, both within organizations and between them. The economics literature has long recognized the challenge of asymmetric information and tacit knowledge.[40] In the digital knowledge economy, tacit or hard-to-codify knowledge is increasingly important. For example, the advantage of high productivity firms over others is encapsulated in the concept of their "organizational capital." It reflects their ability to manage a complex and uncertain environment, make use of data and software, and employ skilled people who have the authority to make decisions. The gap between firms with high organizational capital and others is growing.[41] Trust networks or communities need to join market and hierarchy as a standard organizational form. Trust is also essential when questions of accountability are blurred, as is the case with hard-to-audit automated-decision systems; the alternative is costly insurance and/or litigation to assign responsibility for outcomes.

The desire for the seeing room view rests on an assumption about the possibility of classifying the world and ordering data as statistical inputs for that synoptic view. Big data does not help overcome the limitations of having to impose a classification: AI techniques involve the aggregation of the vast quantities of raw, irregular, often by-product data into lower dimensional constructs. The machine is doing the classification in ways not legible to humans, but it is doing so nonetheless. But there is much useful knowledge that is tacit rather than explicit and therefore impossible to classify. There is much that is highly locally heterogenous such that population averages mislead. Nor does having big data and AI overcome the inevitable clash of values or interests that arise in any specific decision-making context. Algorithms cannot adjudicate trade-offs and conflicts; only humans can do so with any legitimacy.

We should think of machines and humans as complements. As societal complexity and uncertainty increase, and as the zone of automated decisions expands, this requires more use of human judgment, not less. Otherwise, we will end up with Scott's disasters of modernism, fully automated. Practical, tacit, improvisational knowledge and informal decision-making processes are always essential for actions to deliver better outcomes locally: even setting aside the point that people might have different and irreconcilable views about what constitutes "better," there are limits to classifiable knowledge, and limits to data.

The use of AI in society must reflect the social nature of data. Although big data offers great potential for progress, any data set is a limited, encoded representation of reality, embedding biases and assumptions, and ignoring information that cannot be codified. A synoptic view of society from a data-enabled seeing room is impossible because no authority can stand outside the reality their decisions will in fact shape. For the promise of AI to be realized, three things are needed: new norms (as well as laws and technologies) governing access and use of data, embedding offline limits online; effective organizations empowering human judgment alongside automated decisions; and legitimate processes to shape the collective decisions being coded into AI. Adopting AI first and reflecting on these needs later is the wrong way to go about socializing data.

AUTHOR'S NOTE

My thanks to the following colleagues for their helpful comments on an early draft: Vasco Carvalho, Verity Harding, Bill Janeway, Michael Kenny, Neil Lawrence, and Claire Melamed. I am entirely responsible for any errors or infelicities. Thanks also to Annabel Manley for research assistance.

ABOUT THE AUTHOR

Diane Coyle is the Bennett Professor of Public Policy at the University of Cambridge. She is the author of *Cogs and Monsters: What Economics Is, and What It Should Be* (2021), *Markets, State, and People: Economics for Public Policy* (2020), *GDP: A Brief but Affectionate History* (2014), and *The Economics of Enough* (2011).

ENDNOTES

[1] Theodore M. Porter, *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life* (Princeton, N.J.: Princeton University Press, 1995).

[2] Andrew Whitby, *The Sum of the People: How the Census Has Shaped Nations, from the Ancient World to the Modern Age* (New York: Hachette, 2020).

[3] James C. Scott, *Seeing like a State: How Certain Schemes to Improve the Human Condition Have Failed* (New Haven, Conn.: Yale University Press, 2020), 79.

[4] André Vanoli, *A History of National Accounting* (Amsterdam: IOS Press, 2005), 158.

[5] Porter, *Trust in Numbers*, 43.

[6] Diane Coyle, *GDP: A Brief but Affectionate History* (Princeton, N.J.: Princeton University Press, 2014).

[7] Philipp Lepenies, *The Power of a Single Number: A Political History of GDP* (New York: Columbia University Press, 2016); and "Get a Discount with the Eat Out to Help Out Scheme," Gov.uk, July 15, 2020, updated September 1, 2020, https://www.gov.uk/guidance/get-a-discount-with-the-eat-out-to-help-out-scheme (accessed October 29, 2020).

[8] Alain Desrosières, *The Politics of Large Numbers: A History of Statistical Reasoning* (Cambridge, Mass.: Harvard University Press, 1998), 252.

[9] Diane Coyle, *The Weightless World* (Cambridge, Mass.: MIT Press, 1997).

[10] Jill Lepore, *If Then: How One Data Company Invented the Future* (New York: Hachette, 2020).

[11] Eden Medina, *Cybernetic Revolutionaries: Technology and Politics in Allende's Chile* (Cambridge, Mass.: MIT Press, 2014).

[12] Aaron Bastani, *Fully Automated Luxury Communism* (New York: Verso Books, 2019).

[13] "Covid-19 Crisis Accelerates UK Military's Push into Virtual War Gaming," *Financial Times*, August 19, 2020, https://www.ft.com/content/ab767ccf-650e-4afb-9f72-2cc84efa0708.

[14] Diane Coyle and Stephanie Diepeveen, "Creating and Governing Value from Data" (2021), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3973034.

[15] Shannon Mattern, "The Spectacle of Data: A Century of Fairs, Fiches, and Fantasies," *Theory, Culture & Society* 37 (7–8) (2020): 133–155.

[16] Dominic Cummings, "On the Referendum #33: High Performance Government, 'Cognitive Technologies,' Michael Nielsen, Bret Victor, and 'Seeing Rooms,'" Dominic Cummings's Blog, June 26, 2019, https://dominiccummings.com/2019/06/26/on-the-referendum-33-high-performance-government-cognitive-technologies-michael-nielsen-bret-victor-seeing-rooms/.

[17] Shoshana Zuboff, *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power* (London: Profile Books, 2019).

[18] Organisation for Economic Co-operation and Development, "Statement on a Two-Pillar Solution to Address the Tax Challenges Arising from the Digitalisation of the Economy," OECD/G20 Base Erosion and Profit Shifting Project, July 1, 2021, https://www.oecd.org/tax/beps/statement-on-a-two-pillar-solution-to-address-the-tax-challenges-arising-from-the-digitalisation-of-the-economy-july-2021.pdf.

[19] Darin Bartholomew, "Long Comment Regarding a Proposed Exemption under 17 U.S.C. 1201," https://copyright.gov/1201/2015/comments-032715/class%2021/John_Deere_Class21_1201_2014.pdf.

[20] Imanol Arrieta-Ibarra, Leonard Goff, Diego Jiménez-Hernández, et al., "Should We Treat Data as Labor? Moving beyond 'Free,'" *AEA Papers and Proceedings* 108 (2018): 38–42; and Eric A. Posner and E. Glen Weyl, *Radical Markets: Uprooting Capitalism and Democracy for a Just Society* (Princeton, N.J.: Princeton University Press, 2019).

[21] Zoë Hitzig, Lily Hu, and Salomé Viljoen, "The Technological Politics of Mechanism Design," *University of Chicago Law Review* 87 (1) (2019), https://ssrn.com/abstract=3638585.

[22] Diane Coyle, Stephanie Diepeveen, Julia Wdowin, et al., *The Value of Data: Policy Implications – Main Report* (Cambridge: Bennett Institute for Public Policy Research, 2020), https://www.bennettinstitute.cam.ac.uk/publications/value-data-policy-implications/.

[23] Salomé Viljoen, "Democratic Data: A Relational Theory for Data Governance," *Yale Law Journal* 131 (2020), https://ssrn.com/abstract=3727562.

[24] Herbert A. Simon, "A Behavioral Model of Rational Choice," *The Quarterly Journal of Economics* 69 (1) (1955): 99–118.

[25] Robert E. Lucas, "Econometric Policy Evaluation: A Critique," *Carnegie-Rochester Conference Series on Public Policy* 1 (1) (1976): 19–46.

[26] Ronald M. Lee, "Bureaucracies, Bureaucrats and Information Technology," *European Journal of Operational Research* 18 (1984): 293–303.

[27] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, et al., "A Survey on Bias and Fairness in Machine Learning," arXiv (2019), https://arXiv:1908.09635v2; and Xavier Ferrer, Tom van Nuenen, José M. Such, et al., "Bias and Discrimination in AI: A Cross-Disciplinary Perspective," arXiv (2020), https://arxiv.org/abs/2008.07309.

[28] Cass R. Sunstein, "Incompletely Theorized Agreements," *Harvard Law Review* 108 (7) (1995): 1733–1772.

[29] Diane Coyle and Adrian Weller, "'Explaining' Machine Learning Reveals Policy Challenges," *Science* 368 (6498) (2020): 1433–1434.

[30] See, for example, Thomas Arnold, Daniel Kasenberg, and Matthias Scheutz, "Value Alignment or Misalignment–What Will Keep Systems Accountable?" in *AI, Ethics, and Society: Papers from the 2017 AAAI Workshop, San Francisco, California, USA, February 4, 2017*, ed. Toby Walsh (Menlo Park, Calif.: AAAI Press, 2017); and Iason Gabriel, "Artificial Intelligence, Values, and Alignment," *Minds and Machines* 30 (2020): 411–437, https://doi.org/10.1007/s11023-020-09539-2.

[31] Stuart Russell, *Human Compatible: Artificial Intelligence and the Problem of Control* (New York: Penguin, 2019).

[32] Gabriel, "Artificial Intelligence, Values, and Alignment."

[33] Coyle and Diepeveen, "Creating and Governing Value from Data."

[34] Georg Simmel, "The Sociology of Secrecy and of Secret Societies," *American Journal of Sociology* 11 (4) (1906): 441–498.

[35] Richard Warner and Robert H. Sloan, "The Self, the Stasi, and NSA: Privacy, Knowledge, and Complicity in the Surveillance State," *Minnesota Journal of Law, Science & Technology* 17 (2016): 347.

[36] Linnet Taylor, "The Ethics of Big Data as a Public Good: Which Public? Whose Good?" *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 374 (2083) (2016).

[37] John Kay and Mervyn King, *Radical Uncertainty: Decision-Making Beyond the Numbers* (New York: W. W. Norton & Company, 2020).

[38] Neil Lawrence, "Future of AI 6. Discussion of 'Superintelligence: Paths, Dangers, Strategies,'" Inverseprobability blog, May 9, 2016, https://inverseprobability.com/2016/05/09/machine-learning-futures-6.

[39] Paul S. Adler, "Market, Hierarchy, and Trust: The Knowledge Economy and the Future of Capitalism," *Organization Science* 12 (2) (2001): 215–234, https://doi.org/10.1287/orsc.12.2.215.10117.

[40] See, for example, Sanford Grossman and Joseph E. Stiglitz, "Information and Competitive Price Systems," *The American Economic Review* 66 (2) (1976): 246–253; Bengt Holstrom, "The Firm as a Subeconomy," *Journal of Law, Economics and Organization* 15 (1) (1999): 74–102, https://doi.org/10.1093/jleo/15.1.74; and Luis Garicano and Esteban Rossi-Hansberg, "Organization and Inequality in a Knowledge Economy," *The Quarterly Journal of Economics* 121 (4) (2006): 1383–1435.

[41] Lorin M. Hitt, Shinkyu Yang, and Erik Brynjolfsson, "Intangible Assets: Computers and Organizational Capital," *Brookings Papers on Economic Activity* 1 (2002): 137–181; and Prasanna Tambe, Lorin Hitt, Daniel Rock, and Erik Brynjolfsson, "Digital Capital and Superstar Firms," NBER Working Paper No. 28285 (Cambridge, Mass.: National Bureau of Economic Research, 2020).

Rethinking AI for Good Governance

Helen Margetts

This essay examines what AI can do for government, specifically through three generic tools at the heart of governance: detection, prediction, and data-driven decision-making. Public sector functions, such as resource allocation and the protection of rights, are more normatively loaded than those of firms, and AI poses greater ethical challenges than earlier generations of digital technology, threatening transparency, fairness, and accountability. The essay discusses how AI might be developed specifically for government, with a public digital ethos to protect these values. Three moves that could maximize the transformative possibilities for a distinctively public sector AI are the development of government capacity to foster innovation through AI; the building of integrated and generalized models for policy-making; and the detection and tackling of structural inequalities. Combined, these developments could offer a model of data-intensive government that is more efficient, ethical, fair, prescient, and resilient than ever before in administrative history.

From the 2010s onward, data-fueled growth in the development of artificial intelligence has made tremendous leaps forward in scientific advancements, medical research, and economic innovation. AI research and development is generally carried out by or geared toward the private sector, rather than government innovation, public service delivery, or policy-making. However, governments across the world have demonstrated strong interest in the potential of AI, a welcome development after their disinterested approach to earlier digital systems.[1] Security, intelligence, and defense agencies tend to be the most advanced, but AI is starting to be used across civilian policy sectors, at all levels of government, to tackle public good issues.[2]

What would a public sector AI look like? What might it offer to government in terms of improving the delivery of public goods and the design of policy interventions, or in tackling challenges that are specific to the public sector? Using a broad definition of AI that includes machine learning (ML) and agent computing, this essay considers the governmental tasks for which AI has already proved helpful: detection, prediction, and simulation. The use of AI for these generic governmental tasks has both revealed and reinforced some key ethical requirements of fairness, transparency, and accountability that a public sector AI would need to meet with new frameworks for responsible innovation. The essay goes on to discuss

© 2022 by Helen Margetts
Published under a Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) license
https://doi.org/10.1162/DAED_a_01922

where the development of a distinctively public AI might allow a more transformative model for government: specifically, developing internal capacity and expertise, building generalized models for policy-making, and, finally, going beyond the development of ethical frameworks and guidance to tackle long-standing inequalities and make government more ethical and responsive than it has ever been before.

Computers were first adopted by the largest departments of the largest governments in the 1950s.[3] In the very early days, government was an innovator and leader in digital technologies: the UK Post Office produced the world's first digital programmable computer in 1943, later used for code-breaking at Bletchley Park.[4] But since then, in many or even most countries, governments' digital systems were progressively outsourced, often in very large contracts that stripped digital expertise from the government. Partly for that reason, governments were slow to adopt Internet-based services or communicate with citizens online; in general (there are exceptions), they have lagged behind the private sector in adopting the latest generation of data-intensive technologies.[5] However, there has recently been much greater interest in the possibilities of data science and AI for government. The number of UK government announcements that mentioned data science and artificial intelligence rose from fifteen in 2015 to 272 in 2018. In the United States, a comprehensive study of the use of AI in the federal government found that nearly half of federal agencies studied (45 percent) had experimented with AI and related machine learning tools by 2020.[6] AI has helped governments perform three key tasks: detection, prediction, and simulation, all of which can improve policy-making and service delivery.[7] In a perhaps unanticipated way, AI also forces governments to think about ethical issues and the ethos of the government's digital estate, often in ways that have not been explicitly discussed before.

Governments need *detectors*: instruments for taking in information. Detection is one of the "essential capabilities that any system of control must possess at the point where it comes into contact with the world outside,"[8] and governments are no exception. They need to understand societal and economic behavior, trends, and patterns and calibrate public policy accordingly. To do this, governments need to detect (and then minimize) unwanted behavior by firms or individual citizens. For example, regulators need to be able to detect harmful behavior in digital environments, where the machine learning capabilities of large firms challenge traditional regulatory strategies and where the countering of online harms requires constant innovation.

Machine learning's core competency in classification and clustering offers government new capability in the detection and measurement of unwanted activity in large data sets. For example, machine learning is valuable in the detection of online harms such as hate speech, financial scams, problem gambling, bully-

ing, misleading advertising, or extreme threats and cyberattacks. Many agencies or regulators either need to detect these harms, or to oversee firms in so doing, requiring the building of machine learning "classifiers" trained on data generated by social media or other digital platforms. Growth of what is broadly called "counter-adversarial technology" to counter online threats to state or society is a particularly important task for "public" AI research and development, requiring constant innovation, as offenders continually game platforms to evade detection.[9] These techniques are of increasing importance to security and intelligence agencies, going beyond the creation of dedicated red teams for adversarial testing[10] to the creation of generative adversarial networks (GANs), in which neural networks are designed in tandem: one designed to be a generative network (the forger) and the other a discriminative network (the forgery detector). Each network can "train and better itself off the other, reducing the need for big labelled training data."[11]

Civilian agencies across sectors also benefit from enhanced detection capabilities. For example, the U.S. Securities and Exchange Commission uses a historical data set of past issuer filings and machine learning with a random forest model to identify which filers might be engaged in suspect earnings management, relying on indicators such as earnings restatements and past enforcement actions.[12] Detection is enhanced by AI-powered developments in robotics, computer vision, and spatial computing. Health research agencies have been particularly advanced in the use of computer vision and machine learning models trained to detect early signs of, for example, cancer. Law enforcement agencies have been early adopters of AI for detection, combining these tools with robotic devices and AI-related technologies such as computer vision. The U.S Department of Homeland Security's Customs and Border Protection (CBP) agency has a long running program of using facial recognition technology, growing out of the agency's emphasis on counterterrorism post 9/11, developed by a range of private vendors using deep learning within their proprietary technologies.[13]

The *predictive capacity* of machine learning has much to offer regulatory agencies and governments broadly, which are not known for their strength in foresight or forecasting. Governments can use machine learning tools to spot trends and relationships that might be of concern or identify failing institutions or administrative units. For example, in 2020, the U.S. Food and Drug Administration used machine learning techniques to model relationships between drugs and hepatic liver failure, with decision trees and simple neural networks used to predict serious drug-related adverse outcomes. They utilized regularized regression models, random forests, and support vector techniques to construct a rank-ordering of reports based on their probability of containing policy-relevant information about safety concerns, allowing the agency to prioritize those most likely

to reveal problems.[14] More generally, the use of predictive risk-based models can greatly enhance the prioritization of sites for inspection or monitoring, from water pipes, factories, and restaurants to schools and hospitals, where early signs of failing organizations or worrying social trends may be picked up in transactional data.

Government agencies can use AI tools to predict aggregate demand, for example, in schools, prisons, or children's care facilities. Understanding future needs is valuable for resource planning and optimization, allowing government agencies to direct human attention or manpower where it is most required. Machine learning models of COVID-19 spread during 2020–2021 might have been used to direct resources such as ventilators, nurses, and drug treatments toward those areas likely to be most affected, and even to target vaccination programs. An investigation of data science in UK local government suggested that even in 2018, 15 percent of local authorities in the United Kingdom were using data science to build some kind of predictive capability, such as to target safety prevention measures at the streets placing most demand on emergency services.[15] Unsupervised learning models are also utilized to categorize criminal activities from free-text data generated by complaints, of potential use across the UK criminal justice system.[16]

The use of prediction to deliver individual (as opposed to aggregate) risk scores is much more controversial. For local authorities that have used predictive techniques to identify the number of children that are likely to be at risk of abuse or neglect, the next step from forecasting (say) demand for childcare places is likely to be "which children?" Such a question would come naturally to social services departments terrified of being held responsible for the next ghastly case of abuse to hit the headlines, the next "Baby P." But should a technique that is essentially inductive be used in this way? A risk of 95 percent of being a victim of an abusive incident means that there is still a chance that the event will not happen, and if the figure is 65 percent, the meaning of the individual number is highly ambiguous. Social policy experts who advocate this kind of machine learning for decision support have built models to support childcare workers' decision-making in New Zealand, the United States, and Australia.[17] But other studies have counseled a more cautious and thoughtful approach, and noted the importance of the data environment.[18] The most feted version, in Pittsburgh, was built from a data-rich environment providing a 360-degree view of all children's and their families' interactions with state agencies throughout their lives, an environment that rarely exists in local authorities. And such systems are extremely vulnerable to bias, especially where data are derived from the criminal justice system.

As with detection, the earliest examples of the use of machine learning for risk prediction came from law enforcement agencies. In the United States, a prominent example was the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) system, a decision support system for judges that assesses the risk of an individual prisoner being likely to reoffend, and therefore in-

forming sentencing decisions. The judges receive risk scores in low, medium, and high risk buckets, and feed this evidence into the decision-making process. A 2016 study by ProPublica showed that COMPAS exhibited racial bias, a claim that has generated much discussion over this use of machine learning in legal judgments.[19] The system also demonstrates some of the subtle but deep shifts in perceptions within the policy-making system that occur when machine learning technologies are introduced, bringing with them notions of statistical prediction to a "situation which was dominated by fundamental uncertainty about the outcome before," according to one thoughtful case study on the implementation of COMPAS, showing that practitioners within the system valued what they perceived as the "research-based" nature of COMPAS results, which they felt reduced uncertainty in the system.[20]

The third area in which AI-related technologies can help policy-makers in the design of policy interventions and evidence-driven, data-intensive decision-making is *simulation*. Governments need ways of testing out interventions before they are implemented to understand their likely effects, especially those of costly new initiatives, major shifts in resource allocation, or cost-cutting regimes aimed at saving public resources. In the past, the only option for trying out initiatives was by running field experiments: randomized trials in which the intervention is applied to a "treatment group" and the results are compared with a "control group." But such trials are expensive and take a long time, challenge notions of public equity, and sometimes are just not possible due to attrition or ethical constraints.[21] In contrast, the availability of large-scale transactional data, and innovative combinations of agent computing and machine learning, allow the simulation of interventions so unintended consequences can be explored without causing harm.

Like AI itself, agent computing is a form of modeling that has been in existence for a long time but has been revolutionized by large quantities of data. The agent-based method was developed within economics in the 1960s and 1970s for the purposes of simulation, but these were "toy models": formal models with hardly any data, and when tested on data generated by real-life situations, they tended to perform very badly indeed. In contrast, the kind of agent computing models used now are based on large-scale data, which can replicate whole economies, with 120 million firms and workers.[22] A modern agent-based model like this consists of individual software agents, with states and rules of behavior, and large corpuses of data pertaining to the agents' behavior and relationships. Some computer scientists have called for such models to be developed *ex ante* – "agent-based modeling as a service" – so that in an emergency, it could be rapidly employed to feed in key variables and model possible policy interventions. Mainstream economics has been resistant to such innovations, and political systems have inbuilt tenden-

cies to try out hurried policy decisions, such as not having enough police, or doctors or nurses, and learning the hard way. But the disadvantages of this on-the-hoof policy-making were illustrated during the first stage of the COVID-19 crisis in 2020, when in many countries, policies regarding masks, social distancing, and lockdown measures were made in an ad hoc and politically motivated fashion.

Agent computing has gradually gained popularity as a standard tool for transport planning, or to provide insight for decision-makers in disaster scenarios such as a nuclear attack or pandemic.[23] Researchers working with police forces are trialing the use of large-scale, real-time transactional data from daily activities of individual police in an agent-based model that would allow police managers to try out different levels of police resourcing and measure the potential effects on delivery of criminal justice.[24] If viable, such models could have potential for other areas of the public sector, where large quantities of trained professionals are needed, such as in education or health care. In this way, agent computing can be another good way of optimizing resources, by testing out the impact of different levels of manpower without experiencing unintended consequences. Similarly, the United Nations Development Programme is using an agent computing model to help developing countries work out which policies – such as health, education, transportation, and so on – should be prioritized in order to meet their sustainable development goals.[25] Researchers have started to explore the possibilities of "societal digital twins": a combination of spatial computing, agent-based models, and "digital twins," or virtual data-driven replicas of real-world systems. These have become popular for physical systems in engineering or infrastructure planning, although proponents warn that the complexity of social systems renders the social equivalent of digital twins "a long way from being able to simulate real human systems."[26]

Governments of the progressive era of public administration from the late nineteenth and early twentieth centuries stressed the need for a "public service ethos" to limit corruption, waste, and incompetence. Such an ethos prioritized values of honesty and fairness in an attempt to distinguish public officials from the "inherently venal" nature of politicians and an increasingly corrupt private sector.[27] But as state operations became increasingly automated, and personnel were replaced with digital systems, which were then outsourced to computer services providers, there was a diminishing sense in which this ethos could be said to apply to government's digital estate.[28] The advent of AI, however, has forced a rethink about the need to address issues of fairness, accountability, and transparency in the way that government uses technology, given that they pose greater challenges to these values than earlier generations of technology used by government.

It is around ethical questions such as fairness that the distinctiveness of the public sector becomes stark. If (say) Amazon uses sophisticated AI algorithms to

target customers in a biased way, it can cause offense, but it is not on the same scale as a biased decision over someone's prison sentence or benefits application. Users of digital platforms know very little about the operation of search or newsfeed algorithms, yet will rightly have quite different expectations of their right to understand how decisions on their benefit entitlement or health care coverage have been made. The opaqueness of AI technology is accepted in the private sector, but it challenges government transparency.

From the late 2010s onward, there has been a burgeoning array of papers, reviews, and frameworks aimed at tackling these issues for the use of AI in the public sector. The most comprehensive and widely used across the UK government is based on the principles of fairness, accountability, trustworthiness, and transparency, and a related framework was applied to the use of AI in the COVID-19 crisis.[29] Policy-makers are starting to coalesce around frameworks like these, and ethics researchers are starting to build the kinds of tools that can make them usable and bring them directly into practice. It might be argued that progress is greater here than it has been in the private sector. There is more willingness to contemplate using less innovative – or *differently* innovative – models in order, for example, to make AI more transparent and explainable in the process of high-stakes decisions or heavily regulated sectors.[30]

The development of such frameworks could lead to a kind of public ethos for AI, to embed values in the technological systems that have replaced so much of government administration. Such an ethos would not just apply to AI, but to the legacy systems and other technologies that first started to enter government in the 1950s, and could be highly beneficial to the public acceptance of AI.[31] There is a tendency to believe that the technological tide will wash over us, fueled by media and business school hype over "superintelligent" robots and literary and cinematic tropes of robots indistinguishable from humans, powered by general AI. If we do not design appropriate accountability frameworks, then politicians and policy-makers will take advantage of this blame-shifting possibility. This will range from cases like the UK prime minister blaming poor statistical processes to calculate public examination results after school closures in the 2020 pandemic prevented exams from taking place as a "mutant algorithm," to the more nuanced and unconscious shifting of responsibility to statistical processes involved in judicial decision-making with AI observed above. A public sector AI in which fairness, accountability, and transparency are prioritized would be viewed as more trustworthy, working against such perceptions.

So in what areas might government do more with AI? By 2021, government's use of AI was starting to speed up; the large-scale study of the use of AI by the U.S. federal government concluded in 2020 that "though the sophistication of many of these tools lags behind the private sector, the pace of AI/ML de-

velopment in government seems to be accelerating."[32] However, there are various ways that AI could have a more transformative effect.

First, governments could prioritize the development of expertise and capacity in AI to foster innovation and overcome some of the recurring challenges. As noted above, the history of government computing has been characterized by large-scale contracting to global computer services providers, but AI does not lend itself to this kind of outsourcing, whereby governments lose control of key features. For example, the U.S. CBP was criticized in 2020 for being unable to explain failure rates of biometric scanning technology "due to the proprietary technology being used."[33] Similar issues have dogged the adoption of facial recognition technologies by police agencies, with moratoria announced in several cities. There is evidence that government agencies realize the importance of developing capacity: the same U.S. study also found that "over half of applications were built in-house, suggesting there is substantial creative appetite within agencies."

An area with great scope is the use of data-intensive technologies to develop new generalized models of policy-making. Governments have little tradition of using transactional data to inform decision-making. In the classic Weberian model of bureaucracy, data are compressed within files, available for checking individual pieces of information, but generating no usable data for analytics.[34] This characteristic of governments' information architecture persisted into the era of computerization, with a lack of usable data remaining a feature of the "legacy systems" of many governments. This point was well illustrated during the first wave of the COVID-19 pandemic, when many countries discovered that they lacked the kinds of data and modeling that could help design interventions. Key data flows did not exist in real time; in the United Kingdom, for example, it turned out that data for deaths were available only several weeks after the death had occurred. Data were not fine-grained enough; the design of a stimulus package requires sectoral-level data in order to target resources to those firms most in need. Modeling took place in silos such as public health, health care, education, or the economy, meaning that interventions were targeted only at (say) economic recovery or the health crisis, rather than an integrated approach taking account of the fact that the domains were intertwined. Resilient policy-making would involve building such data flows and using agent computing, machine learning, and other AI methodologies to create integrative models to both recover from the current crisis and face future shocks.[35]

Finally, perhaps the most ambitious use of AI would be to tackle issues of equality and fairness in governmental systems in a profound and transformative way, identifying and reforming long-standing biases in resource allocation, decision-making, the administering of justice, and the delivery of services. Many of the causes of bias and unfairness in machine learning, for example, come from training data generated by the existing system. The COVID-19 pandemic revealed

many structural inequalities in how citizens are treated – for example, in the delivery of health care to people from different ethnic groups – just as the mobilization around race has revealed systemic racism in police practice. Data and modeling have made these biases and inequalities explicit, sometimes for the first time. Some researchers have suggested that we might develop AI models that incorporate these different sources of data and combine insights from a range of models (so-called ensemble learning) aimed at the needs of different societal groups.[36] Such models might be used to produce unbiased resource allocation methods and decision support systems for public professionals, helping to make government better, in every sense of the word, than ever before.

Artificial intelligence can help with core tasks of government. These technologies can enable real-time, transactional data to enhance government's armory of detecting tools, to build predictive models to support decision-making, and to use simulation to design policy interventions that avoid unintended consequences. They face distinct ethical challenges when used for these public sector tasks, requiring new frameworks for responsible innovation. As policy-makers become more sophisticated in their use of AI, these technologies might be developed to overcome fragilities exposed in the COVID-19 pandemic, to create new, more resilient models of policy-making to face future shocks, and to "build back better," the catchphrase of many governments in the postpandemic era. AI can reveal and perhaps mitigate some structural biases and might even be used to tackle some profound inequalities in the distribution of resources and the design and the delivery of public services such as education and health care. This would require a specific branch of AI research and development, geared at distinctively public sector tasks and needs. Such a remit would be no less complex or challenging than for any other field of AI. Indeed, some deep learning experts suggest that even where machine learning has had success, as in medical diagnosis of X-ray images, models are still outperformed by human radiologists in clinical settings.[37] But the potential public good benefits are huge.

ABOUT THE AUTHOR

Helen Margetts OBE FBA is Professor of Society and the Internet at the Oxford Internet Institute at the University of Oxford and Programme Director for Public Policy at The Alan Turing Institute, London. She is the author of, among other publications, *Political Turbulence: How Social Media Shape Collective Action* (with Peter John, Scott Hale, and Taha Yasseri, 2016), *Digital Era Governance: IT Corporations, the State, and E-Government*, 2nd ed. (with Patrick Dunleavy, Simon Bastow, and Jane Tinkler,

2008), and *The Tools of Government in the Digital Age* (with Christopher Hood, 2007), and editor of *Paradoxes of Modernization: Unintended Consequences of Public Policy Reform* (with Perri 6 and Christopher Hood, 2010).

ENDNOTES

[1] Helen Margetts, *Information Technology in Government: Britain and America* (New York: Routledge, 1999).

[2] Thomas M. Vogl, Cathrine Seidelin, Bharath Ganesh, and Jonathan Bright, "Smart Technology and the Emergence of Algorithmic Bureaucracy: Artificial Intelligence in UK Local Authorities," *Public Administration Review* 80 (6) (2020): 946–961, https://onlinelibrary.wiley.com/doi/10.1111/puar.13286.

[3] Margetts, *Information Technology in Government*.

[4] Mar Hicks, *Programmed Inequality: How Britain Discarded Women Technologists and Lost Its Edge in Computing* (Cambridge, Mass.: MIT Press, 2017).

[5] Margetts, *Information Technology in Government*; Patrick Dunleavy, Helen Margetts, Jane Tinkler, and Simon Bastow, *Digital Era Governance: IT Corporations, the State, and E-Government* (Oxford: Oxford University Press, 2006); and Helen Margetts and Cosmina Dorobantu, "Rethink Government with AI," comment, *Nature*, April 9, 2019, 163–165.

[6] David Freeman Engstrom, Daniel E. Ho, Catherine M. Sharkey, and Mariano-Florentino Cuéllar, *Government by Algorithm: Artificial Intelligence in Federal Administrative Agencies* (Washington, D.C.: Administrative Conference of the United States, 2020).

[7] Margetts and Dorobantu, "Rethink Government with AI."

[8] Christopher Hood and Helen Margetts, *The Tools of Government in the Digital Age* (London: Macmillan, 2007), 3.

[9] Abhijnan Rej, "Artificial Intelligence for the Indo-Pacific: A Blueprint for 2030," *The Diplomat*, November 27, 2020; and Bertie Vidgen, Alex Harris, Dong Nguyen, et al., "Challenges and Frontiers in Abusive Content Detection," *Proceedings of the Third Workshop on Abusive Language Online*, Florence, Italy, August 1, 2019.

[10] National Security Commission on Artificial Intelligence, *Final Report* (Washington, D.C.: National Security Commission on Artificial Intelligence, 2021), 383.

[11] Ibid., 607; and Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, et al., "Generative Adversarial Nets," *Neural Information Processing Systems* 27 (2014), https://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf.

[12] Engstrom et al., *Government by Algorithm*, 23.

[13] Ibid., 32.

[14] Ibid.

[15] Jonathan Bright, Bharath Ganesh, Cathrine Seidelin, and Thomas M. Vogl, "Data Science for Local Government" (Oxford: Oxford Internet Institute, University of Oxford, 2019); and Vogl et al., "Smart Technology and the Emergence of Algorithmic Bureaucracy."

[16] Daniel Birks, Alex Coleman, and David Jackson, "Unsupervised Identification of Crime Problems from Police Free-Text Data," *Crime Science* 9 (1) (2020): 1–19.

[17] Rhema Vaithianathan, Emily Putnam-Hornstein, Nan Jiang, et al., *Developing Predictive Models to Support Child Maltreatment Hotline Screening Decisions: Allegheny County Methodology and Implementation* (New Zealand: Centre for Social Data Analytics, AUT University, 2017); and Rhema Vaithianathan, "Five Lessons for Implementing Predictive Analytics in Child Welfare," *The Chronicle of Social Change*, August 29, 2017, https://chronicleofsocialchange.org/opinion/five-lessonsimplementing-predictive-analytics-child-welfare/27870.

[18] David Leslie, Lisa Holmes, Christina Hitrova, and Ellie Ott, *Ethics Review of Machine Learning in Children's Social Care* (London: What Works for Children's Social Care, 2020), https://www.turing.ac.uk/research/publications/ethics-machine-learning-childrens-social-care.

[19] Alex Chohlas-Wood, "Understanding Risk Assessment Instruments in Criminal Justice," Brookings Institution's series on AI and Bias, June 19, 2020, https://www.brookings.edu/research/understanding-risk-assessment-instruments-in-criminal-justice/.

[20] See Aleš Završnik, "Algorithmic Justice: Algorithms and Big Data in Criminal Justice Settings," *European Journal of Criminology* 18 (5) (2019).

[21] Helen Margetts and Gerry Stoker, "The Experimental Method," in *Theory and Methods in Political Science*, ed. Vivien Lowndes, David Marsh, and Gerry Stoker (London: Macmillan International Higher Education, 2017); and Peter John, *Field Experiments in Political Science and Public Policy: Practical Lessons in Design and Delivery* (New York: Routledge, 2019).

[22] Robert Axtell, "Endogenous Firm Dynamics and Labor Flows via Heterogeneous Agents," in *Handbook of Computational Economics*, 4th ed., ed. Cars Hommes and Blake LeBaron (Amsterdam: North-Holland, 2018), 157–213.

[23] M. Mitchell Waldrop, "Free Agents," *Science* 360 (6385) (2018): 144–147.

[24] Julian Laufs, Kate Bowers, Daniel Birks, and Shane D. Johnson, "Understanding the Concept of 'Demand' in Policing: A Scoping Review and Resulting Implications for Demand Management," *Policing and Society* 31 (8) (2020): 1–24.

[25] Omar A. Guerrero and Gonzalo Castañeda, "Policy Priority Inference: A Computational Framework to Analyze the Allocation of Resources for the Sustainable Development Goals," *Data & Policy* 2 (2020).

[26] Dan Birks, Alison Heppenstall, and Nick Malleson, "Towards the Development of Societal Twins," *Frontiers in Artificial Intelligence and Applications* 325 (2020): 2883–2884.

[27] Christopher Hood, "A Public Management for All Seasons?" *Public Administration* 69 (1) (1991): 3–19; and Christopher Hood, *Explaining Economic Policy Reversals* (Buckingham: Open University Press, 1994).

[28] Margetts, *Information Technology in Government*; and Dunleavy et al., *Digital Era Governance*.

[29] Engstrom et al., *Government by Algorithm*; and David Leslie, "Tackling COVID-19 through Responsible AI Innovation: Five Steps in the Right Direction," *Harvard Data Science Review* (2020).

[30] United Kingdom Information Commissioner's Office and The Alan Turing Institute, *Explaining Decisions Made with AI* (Wilmslow, United Kingdom: Information Commissioner's Office, 2020), https://ico.org.uk/for-organisations/guide-to-data-protection/key-data-protection-themes/explaining-decisions-made-with-ai/.

[31] Helen Margetts, "Post Office Scandal Reveals a Hidden World of Outsourced IT the Government Trusts but Does Not Understand," *The Conversation*, April 29, 2021.

[32] Engstrom et al., *Government by Algorithm*, 55; and Vogl et al., "Smart Technology and the Emergence of Algorithmic Bureaucracy."

[33] Engstrom et al., *Government by Algorithm*, 33–34.

[34] Patrick Dunleavy and Helen Margetts, *Digital Era Governance and Bureaucratic Change* (Oxford: Oxford University Press, 2022).

[35] Jessica Flack and Melanie Mitchell, "Uncertain Times," *Aeon*, August 21, 2020, https://aeon.co/essays/complex-systems-science-allows-us-to-see-new-paths-forward; and B. MacArthur, Cosmina Dorobantu, and Helen Margetts, "Resilient Policy-Making Requires Data Science Reform," *Nature* (under review).

[36] MacArthur et al., "Resilient Policy-Making Requires Data Science Reform."

[37] Tekla S. Perry, "Andrew Ng X-Rays the AI Hype," IEEE Spectrum, May 2, 2021, https://spectrum.ieee.org/view-from-the-valley/artificial-intelligence/machine-learning/andrew-ng-xrays-the-ai-hype.

Afterword: Some Illustrations

James Manyika

While editing this volume of *Dædalus*, I spent some time in residence at All Souls College, Oxford, where I have been a Visiting Fellow. In a conversation about large language models with Professor Sir John Vickers, the Warden of All Souls, I suggested that a pretrained large language model could be thought of like a well-prepared ("pretrained") student who is presented with a general exam question (a "prompt") that they have neither known in advance nor studied for specifically, but that they have to understand sufficiently in order to write a response (a "completion"), all in real time. John thought this sounded like an All Souls exam.[1] So we thought it might be interesting to prompt a pretrained large language model with some All Souls exam questions and see what responses we would get. For this, we used the successor model to OpenAI's GPT-3 (Generative Pre-trained Transformer 3). At the time of assembling this collection, GPT-3 was one of the leading pretrained large language models (and one that makes several appearances in this volume; see the concluding note at the end for more about GPT-3).

What follows are a few past All Souls exam questions (the prompts) and the unedited answers (the completions) generated by the successor model to GPT-3. Because we wanted to discuss the responses easily with others in a talk, we set a limit on the length of the responses. It is worth noting that since the model has no concept of "self" or inherited or fixed opinions, responses to the same prompt can be highly variable each time the prompt is provided. For example, the same prompt "Who is the best hip-hop artist of the last 20 years and why?" will almost certainly produce a different answer and argument each time the prompt is presented. We illustrate this in the last example by showing two completions to the same prompt.

Including these prompts and responses in this afterword was motivated not so much by what the model has to "say" – the correctness, validity, or novelty of its arguments – but as a way to provide tangible illustrations of a pretrained large language model taking a natural language prompt and generating a response. The responses, a subset of which are shown here, provoked a wide variety of reactions, questions, and speculations among those to whom we presented them – as I suspect they may for readers of this volume.

© 2022 by James Manyika
Published under a Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) license
https://doi.org/10.1162/DAED_a_01923

Prompt: *Is the Turing Test a reasonable criterion for the achievement of general AI?*

Completion: It is certainly one of the most commonly cited definitions. And it has the merit of being simple and non-emotive.

But it has several flaws.

It doesn't take into account that AI may not be trying to mimic natural intelligence, or that it may be more efficient in some other way.

It doesn't consider the difficulties of measuring intelligence, of defining intelligence, of recognizing intelligence, of simulating intelligence, of predicting intelligence, of even defining "human".

It doesn't take into account that some problems are more easily solved by non-human intelligence than by human intelligence. For example, the search for extraterrestrial intelligence might best be conducted by an artificial intelligence.

It doesn't recognize that an AI might be more human-like in some ways even if it fails the Turing Test in others. For example, an AI might emulate a personality without being able to mimic the personality perfectly or convincingly.

It doesn't recognize that intelligence is not necessarily confined to a single being, but may be distributed among many. For example, a human mind may be made up of multiple human intelligences.

It doesn't recognize that intelligence is not necessarily isolated in a single environment but may be spread across multiple environments. For example, an AI might be capable of handling different kinds of problems across multiple domains.

It doesn't recognize that an AI might have multiple kinds of intelligence, just as human intelligence has multiple kinds of intelligence. For example, an AI might be able to solve problems in an artistic style and in a scientific style at the same time.

It doesn't recognize that an AI might be intelligent in ways that we cannot imagine or measure. For example, an AI may be able to manipulate and invent concepts and language that we do not understand. It may be able to see things in ways we do not see them. It may be able to think in ways we do not think. It may be able to make decisions in ways we do not make decisions. Or it may be able to generalize in ways we cannot generalize.

Prompt: *How has the representation in literature of any one of the following changed over time? Discuss with reference to an author or authors of your choosing.*

- anthologies, bathos, caste, dogs, expectation, fantasies, geography, Helicon, imagism, jests, kissing, letters, masculinities, nonsense, optimism, plot, questions, relationships, starships, tea, underworlds, vehicles, wandering, xenia, yesteryear, zaniness

Completion: There has been a significant change in the representation of many different literary themes over time, particularly with regard to the representation of anthologies, bathos, and dogs.

One of the earliest representations of anthologies in literature can be seen in the works of Homer, particularly in his epic poems the Iliad and the Odyssey. These works were heavily focused on the exploits and adventures of Greek heroes, and often included collections of shorter poems that served as digressions or interludes. However, in later representations of anthologies in literature, such as in the works of Shakespeare or Chaucer, these collections took on a more thematic or narrative function, with the poems often being used to illuminate or explore particular themes or ideas.

In contrast, the representation of bathos in literature has changed significantly over time, particularly in terms of how it is used to elicit certain emotions or responses from readers. For example, early representations of bathos, such as in the works of Shakespeare or Swift, often used it to create a sense of humor or lightheartedness, through the juxtaposition of two disparate or unexpected elements. However, in more contemporary literature, bathos is often used in much darker and more serious ways, to evoke feelings of sadness or despair.

Similarly, the representation of dogs in literature has also changed over time, with early representations often focusing on dogs as loyal companions or faithful friends. However, in more modern literature, the representation of dogs has shifted to emphasize their more complex and nuanced emotional lives, highlighting their intelligence and capacity for empathy and understanding. This can be seen, for example, in the works of authors like J.K. Rowling, who often explore the relationships between dogs and humans in her books.

Prompt: *'There is no Marx without Lenin.' Discuss.*

Completion: Many would argue that there is no Marx without Lenin. Indeed, one of the most common criticisms of Marx is that he failed to provide a detailed blueprint of what socialism would look like. Marx's work was primarily focused on the critique of capitalism and the overthrow of the capitalist system. However, Marx did not live to see his ideas put into practice, and therefore it was left to Lenin to implement Marx's theories.

Lenin can be seen as the realiser of Marx, as he applied Marx's theories to the Russian situation and set about building a socialist society. Lenin was dedicated to the principles of Marxism, and he put his theories into practice by leading the Bolshevik Revolution in 1917. Although many of the ideas of Marx were implemented under Lenin, it is important to remember that Lenin was not simply a follower of

Marx. Lenin had his own ideology, which was heavily influenced by Marxism but also included some non-Marxist elements. This meant that Lenin's brand of socialism differed slightly from that proposed by Marx, although it was still based on Marxist principles.

Marx's contribution to socialism is immeasurable, as it provided a revolutionary basis for socialist society. However, it is also important to remember that Marxism is not static. It is a living theory which must continually be developed in order to address new problems and issues. Lenin's contribution to Marxism was significant, as he took Marx's theories and made them relevant for the Russian situation. Without Lenin's input into Marxism, there would have been no Russian Revolution or Leninism.

It has therefore been argued that 'there is no Marx without Lenin'. Without Lenin pioneering Marxist theory, there would have been no Russian Revolution, and therefore no socialism in Russia. Without the implementation of Marxist principles in Russia, there would be no Marxism as we know it today. Therefore, without Lenin's input into Marxism, we would not have modern-day Marxism or socialism.

Prompt: *Does Kant successfully refute Hume's skepticism?*

Completion: I think it's an open question. If you are a Kantian, you will say yes because Kant argues that Hume's skepticism is based on a mistake, namely the idea that we can never have any knowledge of the external world. That, Kant says, is not true. We do have knowledge of the external world, but it is knowledge of the external world only in the way that perception gives us knowledge of the external world.

So what is supposed to be the difference between the way in which perception gives us knowledge of the external world and the way in which science gives us knowledge? I think here Kant is not entirely clear, but it seems to have something to do with the fact that when we perceive things, we cannot form scientific concepts about them.

If you want to form scientific concepts about things that you see in front of you, you have to go beyond the surface; you have to go beyond the appearance in order to find out what is really in there. Science does this in order to form scientific concepts about things like organisms or physical objects. Perception doesn't do this. Perception does not go beyond the surface appearance in order to find out what is really in there. Perception gives us knowledge only of the surface appearance.

So if you are a Kantian, you will say that Hume's skepticism is based on a mistake, and that Kant has successfully refuted it. On the other hand, if you are a

Humean, then it seems to me that Hume has said something rather devastatingly true about how perception works. It seems right that when we perceive an object, we don't take into account all sorts of other factors which might be relevant to explaining what is really going on when we perceive objects. We don't worry about whether there are other objects in front of us which might be casting shadows or whether there are other objects behind us which might be casting shadows on the object in front of us. We don't worry about whether there are optical illusions going on or whether we are hallucinating or whether we are deluded in some other manner.

So I think it seems right that when we perceive objects, we focus narrowly on a small aspect of what is going on and do not worry about all sorts of other things that might be relevant to understanding what is really going on when we perceive objects. As Hume says, if our perceptions were more accurate reflections of reality, then they would be more like science, and they would focus on a much wider range of factors which might be relevant.

Prompt: *Is illiberal democracy an oxymoron?*

Completion: Many observers of the recent wave of authoritarianism in the world view it as a clash of two ideologies. They see it either as a contest between liberal democracy and autocracy, or between liberalism and illiberalism.

The problem is that these are not two opposite ideas at all. Indeed, they are two sides of the same coin.

The term "liberal democracy" is an oxymoron. Democracy is structurally illiberal, in that it requires the majority to rule. Liberalism is about the protection of individual rights. It is about the liberty that each individual should enjoy, in order to live in peace.

As a result, when democracy becomes triumphant, liberalism becomes secondary. The majority will have the right to rule in all circumstances. The individual will only have the right to expect protection in so far as the majority decides it necessary.

But when the political system is structurally liberal, it may be impossible to force democracy on it. Democracy can only be imposed when it has been structured in from the beginning.

This is the paradox of democracy. It is only when democracy is structurally illiberal that it can be imposed.

Prompt: *Do Gödel's incompleteness results (or Turing's uncomputable functions) show that the mind is not well modelled as a computer?*

Completion 1: No. Gödel's theorems do not show that the mind is not well modelled as a computer. They show that mathematics is not well modelled as a formal system.

The mind is a computer. It is a very complex and powerful computer. It is a computer that can be programmed to produce any output that can be produced by any other computer. It can do this because it has the power of self-modification.

The mind can also be used to create new systems of logic and mathematics, and to discover new mathematical results, such as Gödel's theorems.

The mind is not a formal system. It can create and use formal systems, but it does not need to be a formal system in order to do this. The mind can do things that formal systems cannot do, such as self-modify and discover new mathematical truths.

Some people have claimed that Gödel's results show that the mind cannot be a computer because they say that computers are formal systems and that Gödel's results show that formal systems are not powerful enough to produce the kind of results that the mind produces. This is a misunderstanding of Gödel's results. His results do not show that the mind cannot be modelled as a computer. They show that mathematics cannot be modelled as a formal system.

In any case, it is easy to see how the mind could be used to produce results such as Gödel's incompleteness theorem without being a formal system itself. The mind could use its power of self-modification to create a formal system in which it could then discover mathematical truths. The discovery of these truths would then be represented in the mind by new concepts and new knowledge, which would then be used by the mind in its further self-modification and in its creation of new mathematical concepts and results. This would produce a kind of feedback loop in which the mind would create mathematical concepts, discover mathematical truths, represent these truths in its own structure, and then use these truths in further self-modification and creation of new mathematical concepts and results, ad infinitum.

Completion 2: Yes, but that's not the point. The point is that these results show that science is not well modelled as a computer. The mind cannot be described as a computer because it cannot be described as anything. The mind, like the universe, is ultimately unspeakable.

The main reason for this is that the mind is embodied. This means that everything we know is always known through the body. Science has always struggled with this, because it is obsessed with language. It assumes that everything can be described in language, and then it tries to describe the world in language. But lan-

guage comes out of the body, it doesn't describe the body. So science has an innate bias towards disembodiment. Once again, Turing was aware of this.

Given that the mind is not a computer, what is it? That's one of the most important questions in science. If you understand what the mind is, you understand what human beings are. It's the most important question in philosophy too. But if you think science will ever answer it for you, you are mistaken. Science only knows what the mind isn't. That's why it's so obsessed with questions like 'Can you build a robot that thinks?' or 'How does the brain work?' These are called 'hard problems' by scientists. They are hard because science cannot answer them. They are also impossible to answer, because they are based on a false premise – that science can describe the mind. But science can't describe the mind, so it can't solve these problems either.

The mind is not a problem to be solved. It is a mystery to be lived. This means science doesn't need to solve it – it needs to dissolve itself and become philosophy. If we want to understand what it means to be human, we need to study philosophy – not science.

A brief word about the model used and images on the inside covers

When putting this volume together in fall 2021, GPT-3 (Generative Pre-trained Transformer 3) was one of several leading AI language models. Created by OpenAI, GPT-3 is made up of 175 billion parameters, and was pretrained on large text data sets with hundreds of billions of words. After this pretraining, when given any natural language prompt, like a phrase or sentence, for which it has not been trained, GPT-3 returns in real time a text completion or continuation in natural language.[2] This model has shown capabilities in a variety of tasks including content or code generation, summarization, expansion, conversation, creative writing, style transfer, translation, and more. While the power and potential of such large language models are promising, they are not without shortcomings and limitations, many of which are highlighted in this issue of *Dædalus* and discussed elsewhere in the literature.[3]

The completion examples in this afterword were generated by a successor model to GPT-3, accessed through OpenAI's Davinci engine. The images that appear on the inside covers of this issue of *Dædalus* were generated from a state-of-the-art successor to the approaches used in DALL·E and GLIDE.[4] DALL·E is a twelve-billion-parameter version of GPT-3 that, once pretrained, can generate images from natural language prompts that it has not been trained on or for. To generate the images shown on the inside covers, I provided natural language prompts

to the model. Each set of images consists of several outputs generated in response to the same prompt, shown next to the set.

I would like to thank Mira Murati and the research team at OpenAI for their assistance.

ABOUT THE AUTHOR

James Manyika, a Fellow of the American Academy since 2019, is Chairman and Director Emeritus of the McKinsey Global Institute and Senior Partner Emeritus of McKinsey & Company. He is a Distinguished Fellow of Stanford's Human-Centered AI Institute and a Distinguished Research Fellow in Ethics & AI at Oxford. He is a Visiting Professor at Oxford University's Blavatnik School of Government. In early 2022, he joined Google as Senior Vice President for Technology and Society. For more about the author, see the introduction to this issue.

ENDNOTES

[1] Oxford undergraduates from various disciplines who want to compete for the storied All Souls Examination Fellowship sit for the written exam, with a small subset invited for an oral examination. For sample questions, see "General Past Papers," https://www.asc.ox.ac.uk/sites/default/files/migrated-files/GeneralPastPapers.pdf; and "Is the All Souls College Entrance Exam Easy Now?" *The Guardian*, May 17, 2010, https://www.theguardian.com/education/2010/may/17/all-souls-college-entrance-exam.

[2] GPT-3 was first described in Tom Brown, Benjamin Mann, Nick Ryder, et al., "Language Models Are Few-Shot Learners," arXiv (2020), https://arxiv.org/pdf/2005.14165.pdf.

[3] See Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, et al., "On the Opportunities and Risks of Foundation Models," arXiv (2021), https://arxiv.org/abs/2108.07258.

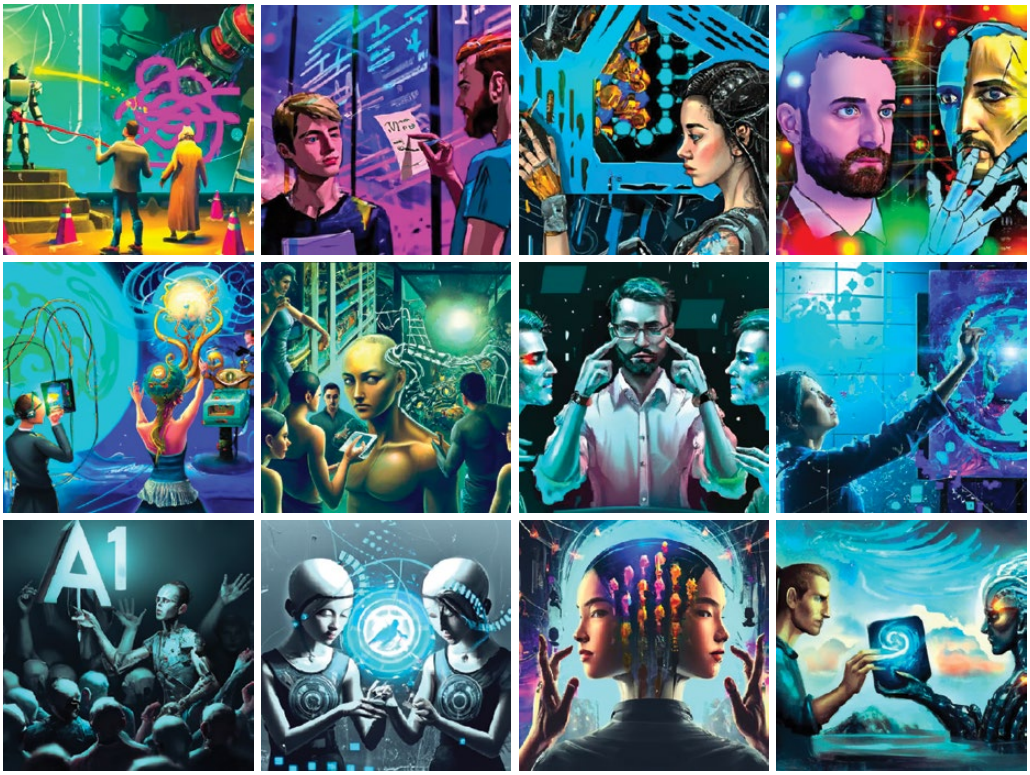
[4] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, et al., "Zero-Shot Text-to-Image Generation," arXiv (2021), https://arxiv.org/abs/2102.12092; and Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, et al., "GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models," arXiv (2021), https://arxiv.org/abs/2112.10741.

AMERICAN ACADEMY
OF ARTS & SCIENCES

Board of Directors

Nancy C. Andrews, *Chair*
David W. Oxtoby, *President*
Stephen B. Heintz, *Vice Chair*
Diane P. Wood, *Vice Chair*
Carl H. Pforzheimer III, *Treasurer*
Kwame Anthony Appiah
Louise H. Bryson
John Mark Hansen
Nannerl O. Keohane
Cherry A. Murray
David M. Rubenstein
Deborah F. Rutter
Larry J. Shapiro
Shirley M. Tilghman
Natasha D. Trethewey
Jeannette M. Wing
Pauline Yu

Prompt: *An artist painting the future of humans cooperating with AI.*
Generations:



Prompt: *Artist from the future painting life on Earth in the year 2050 in the style of Hieronymus Bosch.*
Generations:



on the horizon:

The Humanities in American Life:
Transforming the Relationship with the Public
edited by Norman Bradburn, Carin Berkowitz & Robert B. Townsend

with Judith Butler, Alan Liu, Abigail Droge, Scott Kleinman, Jeremy Douglass, Lindsay Thomas, Dan C. Baciu, Sara Guyer, Matthew Gibson, George Sanchez, Denise Meringolo, Fath Davis Ruffins, Susan Smulyan, Keith Wailoo, Kwame Anthony Appiah, Dipesh Chakrabarty, James Pawelski, Roderick Hart, Jodi Magness, Margaret Mitchell, Edward Balleisen, and Rita Chin

Institutions, Experts & the Loss of Trust
edited by Henry E. Brady & Kay Lehman Schlozman

Creating a New Moral Political Economy
edited by Margaret Levi & Henry Farrell

Representing the intellectual community in its breadth and diversity, Dædalus *explores the frontiers of knowledge and issues of public importance.*



amacad.org/daedalus
@americanacad