EDITED BY

JUSTIN B. BULLOCK

YU-CHE CHEN

JOHANNES HIMMELREICH

VALERIE M. HUDSON

ANTON KORINEK

MATTHEW M. YOUNG

BAOBAO ZHANG

The Oxford Handbook *of* AI GOVERNANCE

OXFORD
UNIVERSITY PRESS

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

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

Library of Congress Cataloging-in-Publication Data
Names: Bullock, Justin B., editor.
Title: The Oxford handbook of AI governance / [edited by] Justin B. Bullock, Yu-Che Chen, Johannes Himmelreich, Valerie M. Hudson, Anton Korinek, Matthew M. Young, Baobao Zhang.
Other titles: Oxford handbook of Artificial intelligence governance
Description: New York : Oxford University Press, 2024. | Includes index. |
Identifiers: LCCN 2023035499 (print) | LCCN 2023035500 (ebook) |
ISBN 9780197579329 (hardback) | ISBN 9780197579343 (epub) |
ISBN 9780197579350
Subjects: LCSH: Artificial intelligence—Law and legislation. |
Artificial intelligence—Moral and ethical aspects. |
Artificial intelligence—Political aspects.
Classification: LCC K564.C6 O94 2024 (print) |
LCC K564.C6 (ebook) | DDC 343.09/99—dc23/eng/20230929
LC record available at https://lccn.loc.gov/2023035499
LC ebook record available at https://lccn.loc.gov/2023035500

DOI: 10.1093/oxfordhb/9780197579329.001.0001

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

Contents

SECTION III: DEVELOPING AN AI GOVERNANCE REGULATORY ECOSYSTEM

VALERIE M. HUDSON

SECTION IV: FRAMEWORKS AND APPROACHES FOR AI GOVERNANCE

YU-CHE CHEN AND MATTHEW M. YOUNG

SECTION V: ASSESSMENT AND IMPLEMENTATION OF AI GOVERNANCE

MATTHEW M. YOUNG AND YU-CHE CHEN

SECTION VI: AI GOVERNANCE FROM THE GROUND UP (VIEWS FROM THE PUBLIC, IMPACTED COMMUNITIES, AND ACTIVISTS WITHIN THE TECH COMMUNITY)

BAOBAO ZHANG

SECTION VII: ECONOMIC DIMENSIONS OF AI GOVERNANCE

ANTON KORINEK

SECTION VIII: DOMESTIC POLICY APPLICATIONS OF AI

JOHANNES HIMMELREICH

SECTION IX: INTERNATIONAL POLITICS AND AI GOVERNANCE

JUSTIN B. BULLOCK

CONTRIBUTORS

Daron Acemoglu, Institute Professor, Department of Economics, MIT

Nakul Aggarwal, MD-PhD Candidate, Medical Scientist Training Program, University of Wisconsin-Madison

Nikita Aggarwal, Lecturer in Law, UCLA School of Law

Anthony Aguirre, Faggin Professor of the Physics of Information, University of California at Santa Cruz

Michael Ahn, Associate Professor, Department of Public Policy and Public Affairs, McCormack Graduate School, University of Massachusetts-Boston

Beatriz Botero Arcila, Assistant Professor of Law, Sciences Po

Amelia C. Arsenault, PhD student, Department of Government, Cornell University, Cornell University

Avital Balwit, Research Scholar, Future of Humanity Institute, Oxford University

Carles Boix, Robert Garrett Professor of Politics and Public Affairs, Princeton University

Joanna J. Bryson, Professor of Ethics and Technology, Centre for Digital Governance, Hertie School, Berlin, Germany

Justin B. Bullock, Associate Professor Affiliate Evans School of Public Policy and Governance, University of Washington

Stephen Caines, Deputy Chief Innovation Officer, City of San Jose

Justin Key Canfil, Assistant Professor, Institute for Politics and Strategy, Carnegie Mellon University

Yu-Che Chen, Isaacson Professor, School of Public Administration, University of Nebraska at Omaha

Jack Clark, Co-Founder of Anthropic; Co-Chair of the AI Index Steering Committee, Stanford Institute for Human-Centered Artificial Intelligence (HAI)

Samuel Clarke, Leverhulme Centre for the Future of Intelligence, University of Cambridge

Cary Coglianese, Edward B. Shils Professor of Law and Political Science, University of Pennsylvania

Allan Dafoe, Senior Staff Research Scientist, DeepMind

David Danks, Professor of Data Science & Philosophy, University of California, San Diego

for AI systems that become increasingly adept at manipulating consumers. Moreover, we need new fairness norms for AI systems that make high-impact decisions that have hitherto been reserved to humans. As the labor market effect of AI and other forms of automation become more severe—and more pernicious for workers—there is also a new need for norms for when and how AI developers should compensate the exposed workers. Even more starkly, if AI makes human workers economically redundant, society will need to establish new norms for how to provide humans with income when labor income is no longer an option (see Korinek & Juelfs, 2024).

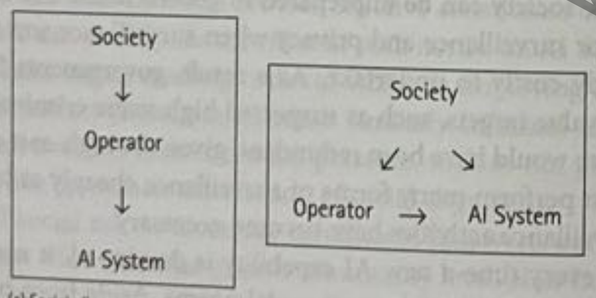
Social Alignment Norms Imposed on Whom?

Up until the recent past, governance to ensure the social alignment of AI systems has relied entirely on society imposing norms on the operators of AI systems, who would be charged with ensuring the alignment of their systems. This is the case, which we may call social alignment by extension, illustrated in panel (a) of Figure 3.2, which employs arrows to indicate that an entity imposes norms on another entity. Such an arrangement would be all that is needed if (1) the operator were perfectly aligned with the social norms, and (2) if the direct alignment between the operator and the AI system held perfectly.

However, when one of these two conditions is violated, it makes it desirable for society to directly impose social norms on AI systems, as illustrated in panel (b) of Figure 3.2. Let us consider each of the two conditions in turn.

When the operator of an AI system is not in compliance with social norms, then imposing social norms directly on AI systems may substitute for the operator's lack of compliance. Such an arrangement may also make it easier to monitor the social alignment of the operator. Consider, for example, an unethical corporation that pursues blind profit maximization to the detriment of other values of society. If the AI systems deployed by such a corporation need to satisfy certain enforceable norms, such as being unbiased, then the space for unethical behavior of the corporation is curtailed. In fact, norms imposed on AI systems may even make it possible to regulate behaviors that violate social norms but were difficult to regulate before. For example, when lending decisions were made individually by loan officers, it was harder to establish whether they were unbiased than it is with algorithms.

When an operator is generally aligned with social norms but has not fully solved the direct alignment problem between her and an AI system, then norms that are imposed directly on the AI system may also help. Such norms can be thought of as best practices, and



(a) Social alignment by extension (b) Social alignment imposed on AI

FIGURE 3.2 Two modes of imposing social alignment norms on AI systems

they may contribute to all three steps of the direct alignment problem that we previously explored: determining the right goal, conveying the goal, and implementing the goal. For example, they may help a well-intended but inexperienced entrepreneur to ensure that the AI system she develops does not unintentionally impose harm on society.

As AI systems become more agentic and have ever more discretion over decisions that used to be reserved for humans, we believe that imposing norms directly on AI systems is becoming increasingly important.[19]

Conclusion

As AI systems become more powerful and are deployed in a growing number of areas, aligning them with our goals becomes ever more vital. However, the expression "our goals" is often used too loosely. It is crucial to emphasize that AI alignment has two distinct dimensions: direct alignment and social alignment. The two dimensions require somewhat different approaches, but we need to solve both to ensure a future that is desirable for humanity. Direct alignment ensures that AI systems pursue goals consistent with the objectives of their operators, irrespective of whether they impose externalities on other parties. By contrast, social alignment ensures that AI systems pursue goals that are consistent with the broader objectives of society, internalizing externalities and considering the welfare of everybody who is impacted by them.

Modern AI systems have the capacity to powerfully optimize for the goals with which we endow them. They are becoming better and better at doing what we are asking them to do and reaching their programmed goals, no matter if these goals correspond to our true goals or if we mistakenly assign them the wrong goals, such as excessively narrow subgoals that lead to disastrous unintended side effects because they fail to fully capture what we want.

Regarding direct alignment, we need to work on determining, conveying, and implementing the goals that we want AI systems to pursue in a robust manner. Regarding social alignment, society needs to determine what social goals and norms we want AI systems to pursue. Social preferences can only determine a partial ordering over all available choices. It is important to expand that ordering as much as possible by resolving social disagreements and conflicts, and to appeal to our better angels as we do this so that our preferences reflect our ethical values. To the extent that society finds agreement, it is also important to develop the right institutions to implement our preferences. We argue that this requires imposing norms on the developers and operators of AI systems as well as new norms that are directly imposed on AI systems.

As AI systems have become more powerful and their use in our world has become more widespread in recent years, we have also witnessed a growing number of cases of social alignment failures, from automated decision systems with biases against disadvantaged groups to social networks that increase polarization and undermine our political systems. Yet progress is continuing, and the powers of our AI systems are continuing to evolve. This makes it urgent to accelerate our efforts to better address the social alignment of AI. If we already have difficulty addressing the AI alignment problems we face now, how can we hope to do so in the future when the powers of our AI systems have advanced by another order

of magnitude? Creating the right governance institutions to address the social AI alignment problem is therefore one of the most urgent challenges of our time.

Acknowledgments

We would like to thank Ondřej Bajgar, Damon Binder, Justin Bullock, Alexis Carlier, Carla Zoe Cremer, Allan Dafoe, Ben Garfinkel, Lewis Hammond, Fin Moorhouse, Luca Righetti, Toby Shevlane, Joseph Stiglitz, and participants at the Spring 2021 Handbook of AI Governance conference for helpful comments and discussions. Any remaining misalignment is our own. We gratefully acknowledge financial support from Center for Innovation, Growth and Society at the Institute for New Economic Thinking (CIGS-INET).

Notes

1. An example is AI-powered financial trading systems (see Boukherouaa & Shabsigh, 2021).
2. Throughout the paper, we use the convention of referring to the entity that is creating, operating, and controlling an AI system as the "operator." In principle, each of these tasks could be performed by different entities, adding additional complexity to the challenge of AI alignment.
3. There are many alternative ways of defining alignment but with similar flavor. For example, an AI system could be aligned to the human's instructions, intentions, revealed preferences, informed preferences, interests, or values, among other options. See Gabriel (2020) for a fuller discussion.
4. This is a shallow definition of agency that is, however, useful for our purposes here. It is inspired by, but distinct from, Dennett's work on stances (see Dennett, 1987). In different contexts, other definitions may be more useful. For example, in ethics, a moral agent is an entity that is morally accountable for its actions. For an elaboration on alternative concepts of agency, see Franklin and Graesser (1996) or Orseau et al. (2018).
5. Another way to express goals is in the form of a "utility function" u(X) that assigns a numerical value to each possible X which ranks the different possibilities. Utility functions are more restrictive than preference orderings. In other words, every utility function defines a set of preferences, but not every set of preferences can be captured by a utility function. For example, for lexicographic preferences, doing so is impossible.
6. Specifically, for a set of preferences to map into a utility function requires several technical assumptions that may be violated, including completeness, transitivity, and continuity.
7. Unfortunately, this nomenclature involves some overloading of the term "agent." In the previous section, we called any entity that can be described as pursuing a goal as an agent; in this section, we follow the conventions of principal-agent theory. Throughout the remainder of this article, the meaning of the term will be clear from the context.
8. In fact, some of the interactions among other social species such as bees or ants can also be described as simple forms of delegation.
9. On the surface, the two described situations—incentivizing a human agent with distinct goals to pursue the principal's goals versus creating an AI agent from scratch who pursues the principal's goals—seem very different. However, given the dualism between actions and goals, there is in fact a deeper equivalence between the two. Addressing the classic principal–agent problem in economics can be viewed as a situation in which the principal has only limited ability to affect the agent's architecture (e.g., to reprogram the agent's primal drive to avoid hard work) and needs to find workarounds ("incentives") to make the agent pursue the desired goal. Programmers frequently experience similar situations. For example, the architecture of ML libraries, say TensorFlow, constrains how they can write their code and makes some results far easier to obtain than others. In other situations, they need to write workarounds building on clunky legacy applications to efficiently obtain the desired behavior. Conversely, human principals sometimes have the ability to "reprogram" agents. For example, parents greatly appreciate the importance of instilling proper goals into their offspring; managers and military leaders know the importance of "inspiring" their agents to pursue desired goals; and a significant part of our human culture (religion, morals, etc.) revolves around reprogramming humans' goals in a way to make our societies operate more harmoniously.
10. For example, Bostrom (2014) describes value alignment as one element of AI control alongside other mechanisms such as capability controls.
11. For a thorough and cutting-edge technical introduction see Russell and Norvig (2020).
12. For example, some definitions of alignment only capture the intentions of the AI system, not the outcome, whether ex ante the AI system was trying to achieve the human's goal. If an AI system tries to accomplish the goal, but some implementation failure causes the system to crash before doing so, perhaps it should not be viewed as a failure of alignment.
13. In mathematics, a full ordering (or total order) is a binary relation on a set that satisfies, among other conditions, that it is transitive and that any two elements are comparable (see https://en.wikipedia.org/wiki/Total_order). The assumption that society's preferences represent a full ordering is also a necessary condition for describing them via social welfare functions.
14. Although it is possible to add such considerations with a negative weight in consequentialist specifications of welfare functions, it is difficult to determine desirable weights.
15. Specifically, when three or more people are asked to express their preferences over three or more alternative choices in pairwise votes, they frequently arrive at outcomes like A > B, B > C and C > A, making it impossible to establish a full order of the available social choices.
16. Even seemingly straightforward mechanisms such as a welfare function that is the sum of utility functions of all members of society will not satisfactorily address the problem, since it may lead to Pareto-dominated outcomes. See Eckersley (2019) for a fuller description of the problem in the context of AI alignment.
17. See https://www.stopkillerrobots.org/.
18. Allowing the harmed individuals themselves to impose a user fee is equivalent to taxing the harm and distributing the revenue to the harmed individuals.
19. In related work, Korinek (2021) proposes the establishment of an AI Control Council to further these objectives.

References

Armstrong, Stuart, Bostrom, Nick, & Shulman, Carl. (2016). Racing to the precipice: A model of artificial intelligence development. *AI & Society* 31(2), 201–206.

Arrow, Kenneth J. (1950). A difficulty in the concept of social welfare. *Journal of Political Economy* 58(4), 328–346.

Bajgar, Ondrej, & Horenovsky, Jan. (2021). Human rights as a basis for long-term AI safety and regulation. Working Paper, University of Oxford.
Baum, Seth D. (2020). Social choice ethics in artificial intelligence. *AI & Society* 35(1), 165–176.
Bessen, James, Impink, Stephen M., Reichensperger, Lydia, & Seamans, Robert. (2021). Ethics and AI startups. *Scholarly Commons at Boston University School of Law*, July 30. https://scholarship.law.bu.edu/faculty_scholarship/1188.
Bostrom, Nick. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford University Press.
Boukherouaa, El Bachir, & Shabsigh, Ghiath. (2021). Powering the Digital Economy: Opportunities and Risks of Artificial Intelligence in Finance. Departmental Paper DP/2021/024, International Monetary Fund.
Christian, Brian. (2020). *The alignment problem*. W.W. Norton.
Christiano, Paul. (2018a). About AI alignment. *AI Alignment*. https://ai-alignment.com/about.
Christiano, Paul. (2018b). Clarifying "AI alignment." *AI Alignment Forum*. https://www.alignmentforum.org/posts/ZeE7EKHTFMBs8eMxn/clarifying-ai-alignment.
Dafoe, Allan, Hughes, Edward, Bachrach, Yoram, Collins, Tantum, McKee, Kevin R., Leibo, Joel Z., Larson, Kate, & Graepel, Thore. (2020). Open problems in cooperative AI. Technical Report, DeepMind.
Dennett, Daniel C. (1987). *The Intentional Stance*. MIT Press.
Eckersley, Peter. (2019). Impossibility and uncertainty theorems in AI value alignment (or why your AGI should not have a utility function). *Proceedings of the AAAI Workshop on Artificial Intelligence Safety*, 1–8. https://dblp.org/rec/conf/aaai/Eckersley19.html?view=bibtex.
European Commission (2016), General Data Protection Regulation (GDPR), Regulation (EU) 2016/679, available at: https://gdpr-info.eu/
Franklin, Stan, & Graesser, Art. (1997). Is it an agent, or just a program? A taxonomy for autonomous agents. In J. P. Müller, M. J. Wooldridge, and N. R. Jennings (Eds.), Intelligent agents III agent theories, architectures, and languages. ATAL 1996. Lecture Notes in Computer Science, vol 1193. Springer, Berlin, Heidelberg. https://doi.org/10.1007/BFb0013570.
Gabriel, Iason. (2020). Artificial intelligence, values, and alignment. *Minds & Machines* 30, 411–437. https://link.springer.com/content/pdf/10.1007/s11023-020-09539-2.pdf.
Hubinger, Evan. (2020). Clarifying inner alignment terminology. *AI Alignment Forum*. https://www.alignmentforum.org/posts/SzecSPYxqRa5GCaSF/clarifying-inner-alignment-terminology.
Human Rights Watch. (2012). Losing humanity: The case against Killer Robots., Technical Report. https://www.hrw.org/report/2012/11/19/losing-humanity/case-against-killer-robots.
Jensen, Michael C., & Meckling, William H. (1976). Theory of the firm: Managerial behavior, agency costs and ownership structure. *Journal of Financial Economics* 3(4), 305–360.
Juechems, Keno, & Summerfield, Christopher. (2019). Where does value come from? *Trends in Cognitive Sciences* 23(10), 836–850.
Kessler, Daniel P. (2010). *Regulation vs. litigation: Perspectives from economics and law*. University of Chicago Press.
Klinova, Katya. (2024). Governing AI to advance shared prosperity. In J. Bullock, B. Zhang, Y.-C. Chen, J. Himmelreich, M. Young, A. Korinek, & V. Hudson (Eds.), *The Oxford handbook of AI governance*. Oxford University Press.
Korinek, Anton. (2021). Why we need a new agency to regulate advanced artificial intelligence: Lessons on AI control from the Facebook Files. Report, Brookings Institution, December 8. https://www.brookings.edu/research/why-we-need-a-new-agency-to-regulate-advanced-artificial-intelligence-lessons-on-ai-control-from-the-facebook-files/.

Korinek, Anton, & Juelfs, Megan. (2024). Preparing for the (non-existent?) future of work. In J. Bullock, B. Zhang, Y.-C. Chen, J. Himmelreich, M. Young, A. Korinek, & V. Hudson (Eds.), *The Oxford handbook of AI governance*. Oxford University Press.
Marquis de Condorcet (1785), Essay on the Application of Analysis to the Probability of Majority Decisions, Paris: Imprimerie Royale.
Ng, Andrew Y., & Russell, Stuart J. (2000) Algorithms for inverse reinforcement learning. *ICML '00: Proceedings of the Seventeenth International Conference on Machine Learning*, 663–670.
Ngo, Richard. (2020). AGI safety from first principles. *AI Alignment Forum*. https://www.alignmentforum.org/s/mzgtmmTKKn5MuCzFJ.
Orseau, Laurent, McGill, Simon McGregor, & Legg, Shane. (2018). Agents and devices: A relative definition of agency. arXiv. https://arxiv.org/abs/1805.12387.
Russell, Stuart J. (2019). *Human compatible: Artificial intelligence and the problem of control*. Viking.
Russell, Stuart J., & Norvig, Peter. (2020). *Artificial intelligence: A modern approach*. 4th US edition. Pearson.
UN General Assembly, Universal Declaration of Human Rights, 10 December 1948, Resolution 217 A (III), available at: https://www.un.org/en/about-us/universal-declaration-of-human-rights
Weber, Max. (1922). Bureaucracy (E. Fischoff, Trans.). Translation of Chapter 6 in *Wirtschaft und Gesellschaft* (pp. 956–1005). Mohr.
Young, Matthew M., Himmelreich, Johannes, Bullock, Justin B., & Kim, Kyoung-Cheol. (2021). Artificial intelligence and administrative evil. *Perspectives on Public Management and Governance* 4(3), 244–258. https://doi.org/10.1093/ppmgov/gvab006.
Yudkowsky, Eliezer. (2004). Coherent extrapolated volition. Machine Intelligence Research Institute. https://intelligence.org/files/CEV.pdf.

CHAPTER 4

THE IMPACT OF ARTIFICIAL INTELLIGENCE

A Historical Perspective

BEN GARFINKEL

Introduction

Over the next several decades, artificial intelligence (AI) is likely to change the world in numerous ways. When trying to describe just how significant these changes could be, commentators tend to reach for historical comparisons. One influential researcher, Andrew Ng, has famously called AI "the new electricity" (Ng, 2017). Elsewhere it is possible to find analogies to fire (Clifford, 2018), nuclear weapons (Allen & Chan, 2017), industrialization (Brynjolfsson & McAfee, 2014), the first computer software (Karpathy, 2021), and even, on occasion, life itself (Tegmark, 2017).

This chapter also takes a history-oriented approach to discussing the impact of AI. Rather than drawing comparisons to individual technologies, however, the chapter instead situates AI within two "reference classes" of technologies that share common traits. If there are any common patterns in how technologies within these reference classes have impacted the world, then we might expect AI to display some of the same patterns.

The first reference class I consider is the set of *general purpose technologies* (GPTs). General purpose technologies are distinguished by their unusually pervasive use, their tendency to spawn complementary innovations, and their large inherent potential for technical improvement. Modern examples include computers, the internal combustion engine, and—in keeping with Ng's suggestion—electricity. Many economists now regard artificial intelligence as an emerging GPT.

I report a few key lessons from the literature on general purpose technologies. One lesson is that the early applications and iterations of GPTs tend to be unassuming. It normally takes several decades for them to achieve large-scale impacts. However, in the long run, a GPT can be expected to alter everything from economic productivity to the character of war to how people spend their leisure time. I attempt to apply these lessons to the specific case of artificial intelligence.

The other, strictly smaller reference class I then consider is the set of *revolutionary technologies*. A revolutionary technology is a GPT that supports an especially fundamental transformation in the nature of economic production. There are only two obvious examples of revolutionary technologies. The first example is domesticated crops, which supported the transition from hunting and gathering to widespread agricultural production. The second example is the steam engine, which supported the transition from an economy where muscle power is the "prime mover" to an economy that is highly mechanized and energy-intensive. Although the concept of a "revolutionary technology" is not a standard one, I believe it is useful for making the point that not all GPTs are created equal. It is plausible that artificial intelligence will eventually emerge as another revolutionary technology by drastically reducing the role of human labor in economic production.

One important lesson from the study of previous revolutionary technologies is that they can facilitate large and long-lasting changes in economic and social trends. For instance, the rate of technological progress increased dramatically around both the Neolithic Revolution and the Industrial Revolution. A number of prominent economists have argued that AI-driven automation could lead to another increase of this sort. If AI does prove to be a revolutionary technology, then it could produce changes that are far more fundamental and far-reaching than anything policymakers have experienced.

General Purpose Technologies in History

General purpose technologies

The concept of a general purpose technology was first developed in the early 1990s by Bresnahan and Trajtenberg (1995). Their central idea was that some technologies simply matter much more than others. As they put it: "Whole eras of technological progress and growth appear to be driven by a few 'General Purpose Technologies' (GPTs)." The key features that distinguish these technologies from others are their unusually pervasive use, their tendency to spawn complementary innovations, and their large inherent potential for technical improvement.[1]

Some evidence for the notion that GPTs have outsized economic impacts comes from the history of total factor productivity (TFP) growth. TFP is a measure of how efficiently investments of labor and other resources can be transformed into goods and services that people wish to buy. It is also often used as a proxy for the rate of technological progress. Historical estimates suggest that, for at least the past century, TFP growth in countries at the economic frontier has come primarily through a small number of waves (fig. 4.1). A common view is that the waves are linked to the widespread adoption of new GPTs (Bresnahan, 2010).

Although Bresnahan and Trajtenberg focused on steam power, electricity, and computers, other authors have since proposed additional technologies as possible GPTs. Table 4.1 includes several of these suggested technologies, drawing from a list generated by Lipsey et al. (2005).

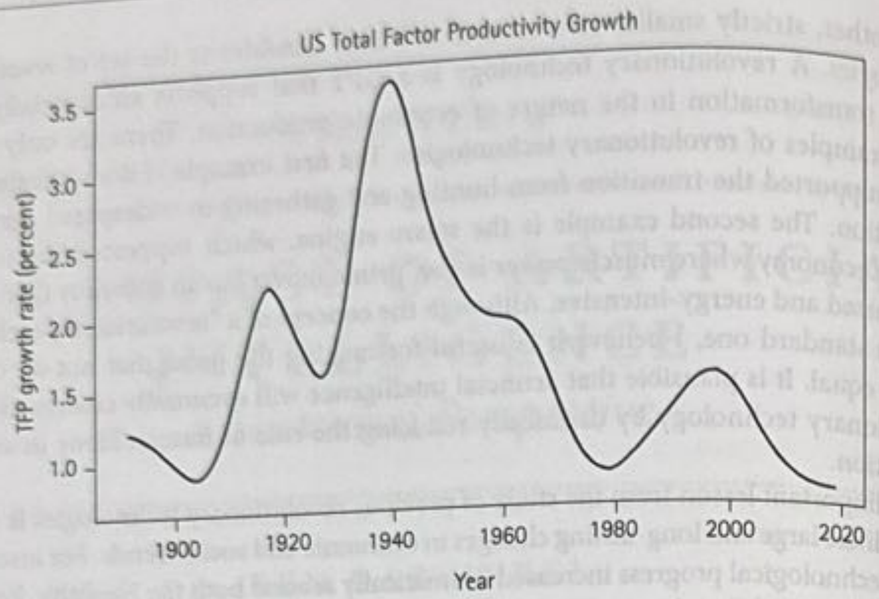


FIGURE 4.1 Waves in the American productivity growth rate, over the past hundred years, according to estimates in the Long-Term Productivity Database (Bergeaud et al., 2017). Note that a high-pass filter, with $\lambda = 500$, has been used to smooth out short-run fluctuations. The most recent wave shown here is normally attributed to the successful adoption of computers and the internet, two recent GPTs (Brynjolfsson & Saunders, 2009). Electrification and the internal combustion engine are often believed to have played outsized roles in the larger, mid-century waves (Bergeaud et al., 2017; Bakker et al., 2019).

Table 4.1 General purpose technologies from across human history, as identified by Richard Lipsey

Technology	First significant use
Domesticated plants	9000–8000 BC
Domesticated animals	8500–7500 BC
Smelting of ore	8000–7000 BC
Wheel	4000–3000 BC
Writing	3400–3200 BC
Bronze	2800 BC
Iron	1200 BC
Water wheel	Early Middle Ages
Three-masted sailing ship	15th Century
Printing	16th Century
Steam engine	18th Century
Railways	19th Century
Internal combustion engine	19th Century
Electricity	19th Century
Computer	20th Century
Internet	20th Century

The concept of a "general purpose technology" is used primarily by economists, with economic impacts tending to serve as the primary standard for inclusion in the category. However, as should be clear from this list, GPTs nearly always have spillover effects on military affairs and politics that are also highly deserving of attention.

The case of electricity

As a useful illustration of the effects a GPT can have across economic, political, and military domains, we can consider the case of electricity. Beginning roughly with the invention of the battery in 1800, or the electric motor in 1821, electricity began to find an increasing array of applications. Beyond enabling methods of long-distance communication like the telegraph, it served as an almost universally applicable method of transmitting energy from the engines or turbines that generated it to machines that could take advantage of it. Although the adoption of electricity proceeded gradually, in countries at the economic frontier, the first half of the 20th century was a period of dramatic "electrification." It became easy to transmit large quantities of energy into individual homes and businesses. Factories were also freed from needing to design their production processes around a single central engine.

We can see the effects of electrification, first, in early twentieth century productivity growth statistics for leading countries such as the United States. We can also see the effects in accounts of how daily life changed for the typical person, as new products like refrigerators, washing machines, lightbulbs, and telephones were introduced (Gordon, 2017). These changes significantly raised living standards, while also helping to reduce the domestic burdens placed on women and very plausibly accelerating their entry into the workforce in the latter half of the twentieth century (Coen-Pirani et al., 2010). Electronic communication technologies like the telegraph and radio also enabled stronger forms of mass political communication and centralized governance, perhaps most notably put to use in a number of mid-twentieth century totalitarian regimes. At the same time, electricity played a key role in enabling a "revolution in military affairs" due to the significant military applications of the radio, the spark-ignition engine, radar, and code-breaking computers (Krepinivich, 1994). The radio, in particular, was core to the *Blitzkrieg* tactics that Germany successfully implemented in the Second World War. In more recent times, electricity has of course enabled all modern information technology and its various impacts as well.

Modest beginnings

The long history of electricity also demonstrates a trajectory common to most GPTs: a GPT typically begins as something crude and only narrowly significant, then slowly achieves a larger impact through decades of technical improvements, costly investments, diffusion of knowledge, individual and institutional adaptation, and further invention.

The history of steam power is perhaps even more remarkable in this regard, given that nearly 200 years passed before it was used for much beyond pumping water out of mines (Von Tunzelmann et al., 1978; Smil, 2018). A similar trajectory has also played out more recently for the computer. In the 1940s, engineers at IBM could apparently see no use for more than a "half-dozen" computers nationwide (Cohen, Welch, Campbell & Campbell, 1999).[2] It was not until the 1990s, about a half-century later, that the introduction of computers had a

noticeable impact on economic productivity, changed most people's daily lives substantially, or were used pervasively in a major military operation (Brynjolfsson & Saunders, 2009).[3]

Revolutionary technologies

If we look back further into the past, then it becomes clear that some periods of technological change are more radical than the rest. Economic historians commonly cite the Neolithic Revolution and the Industrial Revolution as periods that involved unusually fundamental changes to the nature of economic production.[4] I believe it is useful to classify such periods as bringing *revolutionary change*. General purpose technologies that play prominent roles in supporting revolutionary change can then be classified as *revolutionary technologies*.

The Neolithic Revolution involved a transition from an economy largely based on nomadic hunting and gathering to an economy largely based on sedentary agricultural production. It occurred in Western Asia between approximately 10,000 BCE and 5000 BCE, with other regions following after delays of varying lengths. Because domesticated crops played an especially central role in this transition, it is natural to classify them as a revolutionary technology.

The Industrial Revolution involved a transition from an economy largely based on agricultural production to an economy largely based on industrial production and the provision of services. It occurred in Western Europe and the United States between approximately 1750 and 1850, with other regions again following, after delays of varying length. In one interpretation, the Industrial Revolution was an energy transition more than anything else (Landers, 2005; Smil, 2018). The region moved beyond primarily relying on organic sources of energy and material—such as grain, wood, and manure from grass-fed animals—and in doing so opened new productive possibilities. Energy-intensive machines have increasingly supplanted the muscles of humans and animals. Because the steam engine played a very important role in this transition, it can also be classified as a "revolutionary technology."[5]

Shifts in the human trajectory

Both the Neolithic Revolution and Industrial Revolution were accompanied by a number of dramatic and long-lasting changes in economic and social trends. In other words, beyond their qualitative impact on economic production, both these revolutions can be said to have shifted the human trajectory in significant ways.

Energy capture

Perhaps the most fundamental trend to emphasize is growth in "energy capture," a measure of the total amount of energy harnessed by humanity (Morris, 2013). This includes energy captured through food eaten by humans and domesticated animals, through fuel used to produce heat or power machines, and through the accumulation and alteration of physical materials. For obvious reasons, energy capture is associated with the capacity to support a large population, wage war, manufacture goods, travel, process information, and generally,

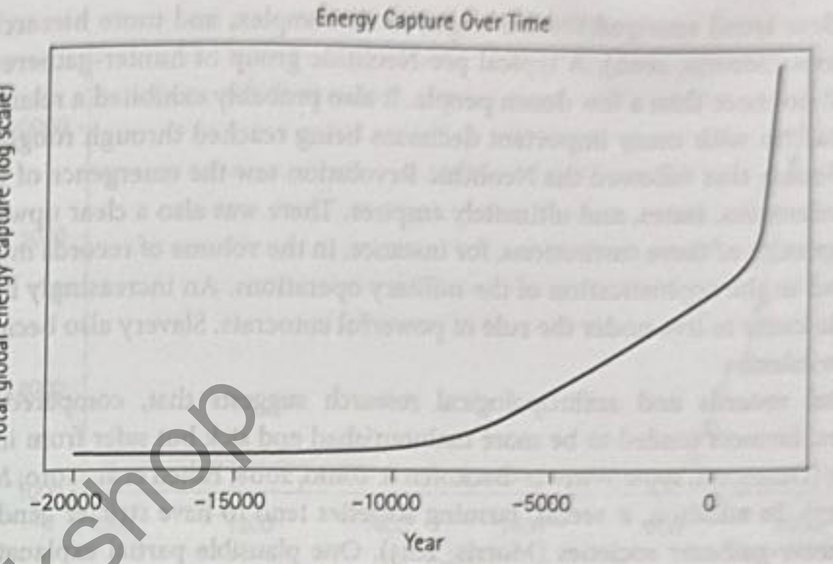


FIGURE 4.2 A stylized depiction of global energy capture over time. Although we lack reliable numerical estimates of the growth rate before the twentieth century, we can be fairly confident that it increased dramatically around both the Neolithic Revolution and the Industrial Revolution.

as the historian Ian Morris puts it, "get things done in the world" (Morris, 2010). Higher rates of energy capture growth have tended to reflect higher rates of technological progress and material change. Energy capture also lends itself more naturally to discussions of long-run trends than more sophisticated economic metrics. For instance, metrics such as gross domestic product and total factor productivity are not obviously applicable to hunter-gatherer societies.

Unsurprisingly, there are no reliable numerical estimates of global energy capture in pre-modern times. Nonetheless, we can be fairly confident that the rate of growth increased dramatically over the course of both economic revolutions (Smil, 2018). Sedentary farming societies, which manage dense concentrations of high-value plants and animals, extract far more energy per unit of land than mobile hunter-gatherer societies do. As a result, the spread and intensification of agriculture over several thousand years likely enabled an unprecedented rate of energy growth. The later transition away from organic sources of energy and toward fossil fuels, which began around the Industrial Revolution, then raised the rate of energy growth to even greater heights.[6]

Figure 4.2 presents a stylized depiction of global energy capture over time.[7] The events of the past century have all unfolded in the context of the nearly vertical section of the graph. When we looked at recent waves of technological change, this obscured the fact that even the lacunas between the waves contained unusually rapid change by pre-industrial standards. Ever since the Industrial Revolution, humanity's ability to "get things done in the world" has been growing at a rather exceptional pace.

Further shifts: The Neolithic Revolution

Most trend changes associated with the Neolithic Revolution are, unsurprisingly, fairly uncertain. Nonetheless, as groups of humans settled down and became farmers,

a very clear trend emerged toward larger, more complex, and more hierarchical social institutions (Morris, 2010). A typical pre-Neolithic group of hunter-gatherers probably included no more than a few dozen people. It also probably exhibited a relatively flat social hierarchy, with many important decisions being reached through rough consensus. The millennia that followed the Neolithic Revolution saw the emergence of increasingly large settlements, states, and ultimately empires. There was also a clear upward trend in the complexity of these institutions, for instance, in the volume of records maintained by states and in the sophistication of the military operations. An increasingly large portion of people came to live under the rule of powerful autocrats. Slavery also became increasingly prevalent.

Skeletal records and anthropological research suggests that, compared to hunter-gatherers, farmers tended to be more malnourished and sick but safer from interpersonal violence (Diamond, 1998; Wittwer-Backofen & Tomo, 2008; Eshed et al., 2010; Morris, 2014; Gat, 2017). In addition, it seems, farming societies tend to have stricter gender divisions than hunter-gatherer societies (Morris, 2015). One plausible partial explanation for this sharpening of gender roles is that the transition to farming increases the importance of upper body strength, for work outside the home, and men typically have more upper body strength than women. At the same time, women tend to become more tied to their homes, partly because sedentism allows for larger family sizes. Therefore, as agricultural practices diffused across the world, it is fairly likely that average living standards, levels of gender equality, and levels of violence all declined over thousands of years.

Further shifts: The Industrial Revolution

Perhaps the most notable outcome of the Industrial Revolution was the beginning of sustained growth in the average person's wealth (fig. 4.3). The current tendency for each generation to be noticeably wealthier than the generation that came before is historically anomalous. The transition to a much higher energy capture growth rate certainly played an important role in supporting the emergence and sustainability of this trend. Per-capita growth is only possible when total output grows quickly enough to outstrip the population growth rate.

The Industrial Revolution was also accompanied by a new trend toward more widespread democracy. Before the eighteenth century, democracy was an unusual form of government. Although some pre-modern states did have democratic elements, such as assemblies that constrained the actions of rulers, the ability of common people to participate was typically quite limited.[8] Furthermore, at the start of the eighteenth century, there was no clear sign of a global trend toward widespread democracy. Nonetheless, over the past 200 years, an obvious trend has emerged (fig. 4.4).

The Industrial Revolution was certainly not the sole cause of this new trend (Stasavage, 2020). A number of seemingly critical intellectual and political developments either predate the revolution or appear rather disconnected from it. The American Revolution, for instance, cannot be chalked up to industrialization. Nonetheless, many economists and historians do contend that the Industrial Revolution is an important part of the overall story. For instance, Acemoglu and Robinson (2006) suggest that agrarian economies are naturally less likely to democratize. So long as elites derive most of their income from rents on large tracts of land, they will have an especially strong reason to resist universal suffrage.

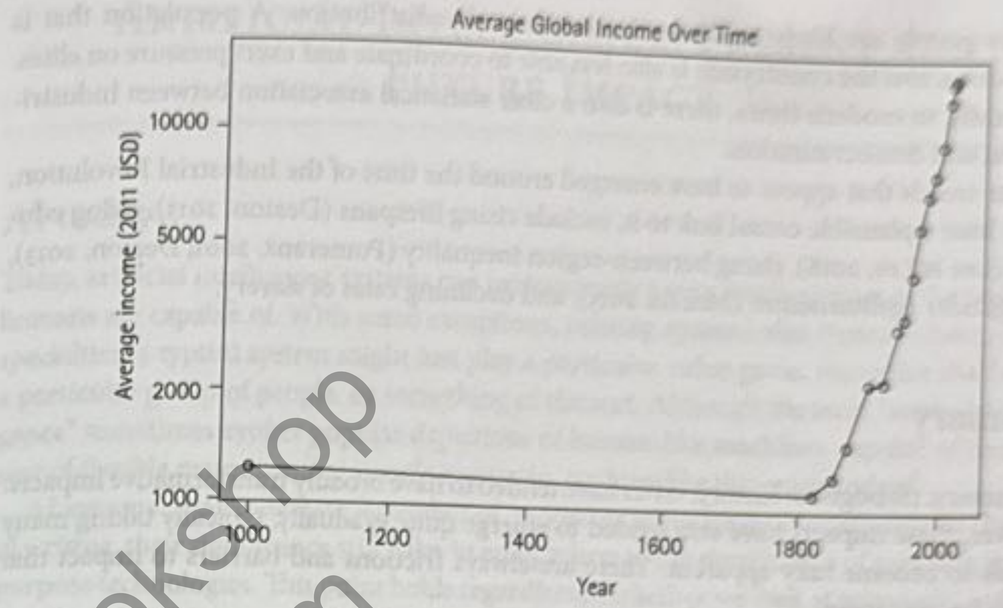


FIGURE 4.3 The average global income over time, according to estimates in the Maddison Project Database (Bolt & Van Zanden, 2020). The data point for 1000 CE is based solely on an estimate for China, which contained a large portion of the world's population at the time. Although the specific numbers given in the dataset are controversial, the qualitative story they suggest is not. There was no substantial, sustained, global income growth before the Industrial Revolution.

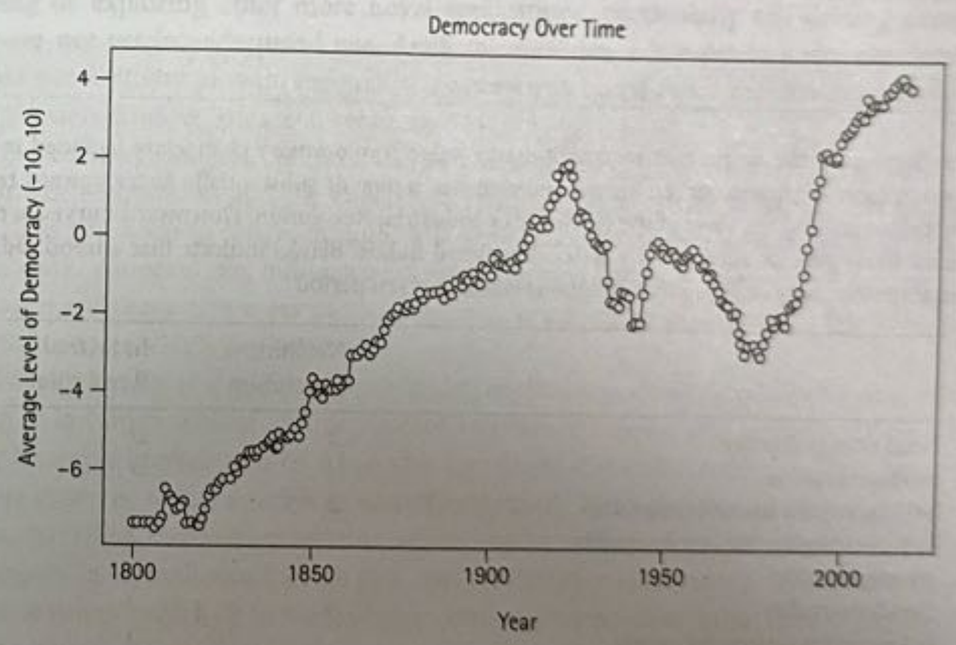


FIGURE 4.4 According to estimates in the PolityV dataset, the average level of democracy has been increasing for the past two hundred years (Marshall & Gurr, 2020).

ordinary people are likely to demand massive land redistribution. A population that is spread out across the countryside is also less able to coordinate and exert pressure on elites. Empirically, in modern times, there is also a clear statistical association between industrialization and democratization.

Other trends that appear to have emerged around the time of the Industrial Revolution, with at least a plausible causal link to it, include rising lifespans (Deaton, 2013), rising education (Lee & Lee, 2016), rising between-region inequality (Pomeranz, 2001; Deaton, 2013), rising gender egalitarianism (Morris, 2015), and declining rates of slavery.

Summary

In summary, throughout history, GPTs have tended to have broadly transformative impacts. However, these impacts have also tended to emerge quite gradually, typically taking many decades to become fully apparent. There are always frictions and barriers to impact that need to be overcome.

At least two GPTs, domesticated crops and the steam engine, stand out from the rest. These two *revolutionary technologies* both supported unusually fundamental changes to the nature of economic production. These economic changes in turn supported unusually significant and long-lasting shifts in the trajectory of humanity (Box 4.1).

When considering the future of artificial intelligence, it is useful to ask whether it will constitute a general purpose technology. We should also ask, though, whether it has the potential to become something even more transformative. If so, then popular analogies to recent GPTs such as electricity and the internal combustion engine may actually understate its ultimate significance.

Box 4.1

Some features of the world that we might use to judge revolutionary change are included in the following table. Upward-curved arrows indicate that a new or substantially faster upward trend likely began around the time of the Neolithic or Industrial Revolution. Downward-curved arrows indicate likely new or substantially faster downward trends. Blanks indicate that a trend did not change substantially or is highly ambiguous within a given period.

	Neolithic Revolution	Industrial Revolution
Total energy capture	↗	↗
Average income		↗
Within-region income inequality	↗	
Between-region income inequality		↗
Average health	↘	↗
Gender equality	↘	↗
Information processing power		↗
War-making capacity	↗	↗
Collective decision-making and political freedom	↘	↗
Degree of political centralization	↗	↗

Artificial Intelligence: Present and Future Impact

AI today

Today, artificial intelligence systems can perform only a very small portion of the tasks that humans are capable of. With some exceptions, existing systems also typically have narrow specialties: a typical system might just play a particular video game, recognize the faces of a particular group of people, or something of the sort. Although the term "artificial intelligence" sometimes evokes popular depictions of human-like machines, capable of the same sort of flexible reasoning that people engage in, nothing like this exists today.[9]

AI systems do already have a number of important applications. Nonetheless, at the time of writing, their significance still pales in comparison to the significance of previous general purpose technologies. This point holds regardless of whether we look at economic, military, or political impacts.

Economically, AI's most profitable present-day and near-term applications appear to be improved online marketing and sales systems, which make recommendations, select offers, and target advertisements in response to user data. Systems that help with supply chain optimization, for instance, by suggesting changes to order sizes and schedules, also appear to be highly valuable.[10] Other applications include more accurate fraud detection and more efficient data center cooling. While there are many other miscellaneous applications, such as the speech recognition software now installed on most smartphones, their total value to consumers still appears to be modest. Large investments have also been made into developing or exploring other more novel applications, particularly self-driving cars, but these are not yet in widespread use. Artificial intelligence has yet to have any clear impact on productivity growth, inequality, unemployment, or other macroeconomic trends (Brynjolfsson, Rock & Saunders, 2019).

In the military domain, its most valuable present-day applications may be associated with image recognition and the analysis of bulk-collected reconnaissance data (Pellerin, 2017). It is too soon to judge, though, just how much advantage will be gained through recent work. Autonomous military vehicles and much better systems for detecting cyber intrusions are also active areas of investigation, in a number of countries, but have yet to materialize or be widely deployed.

Politically, the most significant present-day applications of artificial intelligence may be related to law enforcement, online content recommendations, and target political advertising. Example applications of AI in the domain of law enforcement include predictive policing systems, which attempt to identify especially likely locations and perpetrators of crimes, and facial recognition systems, which can be used to identify and track individuals who appear in surveillance footage (Joh, 2017; Whittaker et al., 2018). A common hope is that these systems will help to reduce crime, while a common fear is that they will exacerbate or make it harder to reduce existing intergroup disparities in treatment by law enforcement. Some commentators worry that AI systems that recommend Facebook groups, YouTube videos, or other digital content based on users' behavior could have the effect of increasing political polarization or promoting false beliefs (Harris, 2020). A related concern is that

some targeted political advertisements could be substantially more effective or manipulative than typical political advertisements. Existing systems for generating convincing fake photographs and videos could also become politically significant, for instance if they are used to generate false depictions of political figures in compromising situations (Brundage et al., 2018). Again, though, most systems in this category are either not yet in widespread use or not yet effective enough to have had obvious society-level impacts on crime rates, voting patterns, intergroup disparities, levels of incarceration, and the like.[11]

Barriers to impact

Overall, the impact of AI has been limited in two ways. First, there are *technical bottlenecks* on what AI engineers can accomplish today. This means that existing techniques, sources of data, and quantities of available computing power are not sufficient to develop AI systems capable of performing many tasks of interest. The tasks that AI systems can perform today, such as targeted advertising and image recognition, possess a number of somewhat unusual traits that make them especially tractable.[12]

Second, there are *implementation challenges*. This means that even for potential applications that are not currently "out of bounds," the process of discovering, developing, and widely deploying these applications may be quite slow. Limiting factors include the scarcity of expertise, regulatory barriers, the unavoidable complexity of many engineering projects, the need to invent complementary technologies and services, and the need to attract large investments of capital. The relatively slow process of getting self-driving cars into widespread use has provided a clear illustration of several of these factors (Fagella, 2020).

AI's general purpose potential

Although its impact remains comparatively modest, artificial intelligence is a promising candidate for a new general purpose technology. As we have seen, it is already being applied in a wide range of domains. It is also inspiring enormous research and development efforts, which, by some estimates, might now account for substantially more than one percent of the world's total R&D spending.[13] Furthermore, if researchers can make enough progress on relevant technical bottlenecks, then the technology's potential for long-term improvement is enormous.

In fact, a growing number of economists have begun to identify AI as a likely GPT. This includes Manuel Trajtenberg, the founder of the GPT literature, and Erik Brynjolfsson, who is also a leading expert on the economic impact of information technologies (Trajtenberg, 2019; Brynjolfsson, 2019).

As discussed above, almost every GPT that has been developed so far began as something extremely crude with only a handful of practical applications. Artificial intelligence could then be in the early stages of an impact trajectory that several other technologies have followed before. In keeping with the two varieties of limitations mentioned above, there are roughly two reasons we might expect the impact of AI to grow over time. First, we might expect technical bottlenecks to decrease significantly over time, thereby "unlocking" many new applications.[14] Second, even without much of a change in capabilities, we can

reasonably expect many more applications to be developed and come into widespread use in the coming decades. The process of discovering what is already possible and implementing it could continue for a very long time.[15]

Economically, two especially significant applications that might be nearly within reach are self-driving cars and customer service systems. Given that over five million Americans are currently employed as vehicle operators or in call centers, Erik Brynjolfsson argues that automating a large portion of these roles over the next few decades would significantly boost national productivity and impact many workers' lives (Brynjolfsson, 2019). There has also been major recent progress in developing various kinds of *generative models*, such as systems that produce illustrations when given text prompts, that produce essays when given opening sentences, and even systems that produce lines of code when given descriptions of the desired code (Brown et al., 2020; Chen et al., 2021). At the time of writing, these systems are not yet ready for very widespread commercial use. However, progress continues to be rapid, and highly valuable applications might ultimately be very close at hand. A final promising application area to note is biomedical research. A recent major breakthrough in protein structure prediction suggests that artificial intelligence could find significant near-term applications related to drug design (Jumper et al., 2021).

Various authors have produced dramatic estimates of the portion of current jobs that could ultimately be automated, given present capabilities, often producing numbers in the double digits (Arntz et al., 2016; Chui et al., 2016; Frey & Osbourne, 2017; Winick, 2018). Some economists have also suggested, controversially, that such a wave of automation drawing on existing AI techniques could raise income inequality or unemployment.[16] However, there is still no consensus about the near-term economic impact of AI (Cukier, 2018).

In the military domain, present capabilities may be sufficient to develop greatly improved autonomous vehicles, systems for analyzing reconnaissance data, and systems to aid cyber offense and defense. Weaponized drone swarms, intended to overwhelm the defenses of large weapons platforms such as aircraft carriers, are one application that could have an especially large impact on the character of war (Scharre, 2014). There may also be valuable applications involving the acceleration or improvement of behind-the-scenes processes, such as vetting individuals for security clearances or determining vehicle maintenance schedules (DARPA, 2018). Ultimately, artificial intelligence could increase military effectiveness substantially. It might also shift important strategic parameters, such as the likelihood of accidental escalation, the relative ease of offense and defense, or the speed of military power transitions (Allen & Chan, 2017; Horowitz, Allen, Kania & Scharre, 2018; Scharre, 2018). Such shifts might then have an influence on the likelihood of war.

In the political domain, we might see continued improvement and adoption of systems for more effective law enforcement, political persuasion, and video forgery. Some authors have raised concerns that these applications could make open, fact-based political discourse more difficult and bolster authoritarian regimes (Brundage et al., 2018). More positive effects might include reduced crime or risk from terrorism due to law enforcement applications, or improved government decision-making due to better data-driven analysis. The possibility of increased inequality, unemployment, or international economic and military competition could also pose substantial political challenges (Dafoe, 2018; Horowitz et al., 2018).

Table 4.2 summarizes some existing and potential applications, including the several just discussed. Arguably, this list is still substantially less radical than the list of applications