MERCATUS SPECIAL STUDY

ARTIFICIAL INTELLIGENCE

AN INTRODUCTION FOR POLICYMAKERS

REVISED EDITION

Matthew Mittelsteadt, ***Mercatus Center***

MERCATUS.ORG

MERCATUS CENTER
George Mason University

Matthew Mittelsteadt, *Artificial Intelligence: An Introduction for Policymakers*, rev. ed. (Mercatus Special Study, Mercatus Center at George Mason University, October 2024).

Abstract

This introduction seeks to equip a diversity of policymakers with the core concepts needed to identify, understand, and solve artificial intelligence (AI) policy challenges. AI is best conceived as an often ill-defined goal, not a monolithic general-purpose technology, driven by a diverse and ever-evolving constellation of input technologies. The document first introduces a sample of AI-related challenges to ground the importance of understanding this technology, the diversity of issues it will create, and its potential to transform law and policy. Next it introduces AI, key terms such as machine learning, and ways that AI progress can be assessed. Finally, it introduces and explains how three key input technologies—data, microchips, and algorithms—work and make AI possible. These core technologies are known as the AI triad. Intended to serve a variety of audiences, these explanations are presented with multiple levels of depth. Technical concepts are tied to relevant policy questions, thereby guiding the application of this knowledge while illustrating the value of understanding this emerging technology beyond a surface level. This introduction to AI appears both in written form and as an ever-evolving website supported by the Mercatus Center.

JEL codes: O38, O30, O31, O32, O33, C63

Keywords: Artificial intelligence, AI, machine learning, ML, neural network, technology, science, deep learning, intelligence, reinforcement learning, AI policy, emerging technology, algorithms, data, big data, semiconductors, microchips, chips, policy, law, autonomous systems, autonomy, LLMs, models, technology policy, primer, computer science, computation, prediction, engineering, robotics, computation, general purpose technology, GPT, public administration

Contents

1. Introduction

In 2022, the release of ChatGPT took both the public and policymakers by surprise. Accessible, conversationally fluid, and helpful across a breadth of domains, **artificial intelligence (AI)** finally started to distantly resemble sci-fi promises. We are clearly at an AI inflection point, but what has changed? First is scale. Near the turn of the current decade, three forces converged to enable large, massively intelligent models: the flexibility of new, highly scalable algorithmic techniques, big data amassed at scale to provide necessary knowledge, and the wickedly fast chips to handle AI's highly intensive data and computation. Second, and perhaps more important, is breadth. In 2021, Stanford University's Human-Centered Artificial Intelligence Institute wrote about a broad, profound AI "paradigm shift" arising from these convergent technologies.[1] The result of AI's new scale was a new class of so-called **foundation models**—large-scale systems trained on broad sets of data that can be easily adapted or **fine-tuned** to a wide range of downstream tasks.[2] Armed with computational heft and flexibility, these models offer many of the tools needed for AI to step beyond a mere curiosity.

Even with a measure of skepticism toward the AI hype, recent developments illustrate that the impact of AI is simultaneously emerging across a multitude of domains:

- AlphaFold predicts the structure of nearly every known protein—a critical new tool in biological and medical research.
- Midjourney's art generator produced near-human-quality works.
- AlphaTensor discovered a more efficient approach to matrix multiplication than previously known, and this could soon speed up a wide range of applications.
- Insilico Medicine's proprietary systems created a potential treatment for idiopathic pulmonary fibrosis, the first AI drug to enter FDA phase II human trials.[3]

- MūtCompute discovered an enzyme that breaks down polyethylene terephthalate, a common plastic that represents 12 percent of global waste.
- GraphCast provides 10-day weather predictions with state-of-the-art accuracy in under a minute.[4]
- OpenAI's ChatGPT produces logically complete text and image responses to user queries.[5]

Note the range of fields. AI is being applied everywhere—from the arts and linguistics to chemistry and pure mathematics. It is flexible. The tools that make AI possible represent a new class of **general-purpose technologies**, innovations that "[have] the potential to affect the entire economic system."[6] Just as previous general-purpose technologies such as electricity transformed society, AI systems are changing many domains—from science to entertainment, from education to health, from national defense to the financial system—and could even radically transform them.

Some critics claim that these advances in AI are skin deep, mere "**stochastic parrots**" that randomly rearrange and regurgitate data.[7] They may look effective, critics argue, but AI lacks any true understanding, common sense, or ability to explain its decisions. There is ample room for debating the nature of true intelligence, and critics will err dramatically if they dismiss AI outright as unimportant. The future of AI is unclear, but the increasing breadth and scale of AI applications demands attention from an increasing breadth of decision makers.

Yet policymakers are often not keeping pace.

Policy decisions about AI made today may hold long-term importance for this technology's future. While knowledge is no "good government cure-all," it is a necessary first step for thoughtful decision-making. In this work, we hope to impart a basic understanding of AI design, application, and policy challenges to inform policy-minded readers.

The Tip of the AI Policy Iceberg

"However brilliant computer engineers may be when facing down technological challenges, they rarely have real insight into what's happening outside the digital bubble."[8]

—Jacob Helberg, former Google news policy lead; commissioner, US-China Economic and Security Review Commission

What do we lose without a diversity of experts engaging with AI in depth?

In summer 2022, AI image generation seemed to appear out of nowhere. With the release of DALL-E mini, an open-source approximation of OpenAI's DALL-E 2 art generator, AI art was suddenly accessible to everyone. Delighted by the often strange yet sometimes human-quality works, consumers flocked to the application and flooded social media with bizarre AI creations. Powerful enough to wow yet amusingly inaccurate, DALL-E mini introduced many to a glimpse of the possibilities of image generation while comforting others with the understanding that **generative AI** was still out of immediate reach. Yet, in just a matter of weeks, things changed. OpenAI broadened access to the full version of DALL-E 2, Midjourney generated covers for *The Economist*,[9] and Stability AI released the powerful Stable Diffusion— in just one summer these wonky generators suddenly proved capable. Since then, generative tech has matured and broadened, perfecting images while making headway in video, audio, and video games.

This sudden burst of innovation caught policy officials off guard. In a matter of weeks, copyright, intellectual property (IP), and other media-relevant officials had to shift gears toward confronting an unexpected slew of novel AI-based issues far from consideration just months earlier. One notable controversy: artistic rights. To develop these systems, engineers scraped volumes of preexisting human-crafted works from galleries across the web, leaning on those data to hone their model's image-crafting abilities. Often, this process was undertaken without artistic consent. As a result, prominent digital artists found that this software could produce near-perfect renditions of their works, allowing anyone to appropriate signature styles.[10] This situation raised challenging questions of usage rights, privacy, personal autonomy, and copyright infringement. Since 2022, the controversy has led to growing artistic agitation. Online, creators have attempted to disable image generators through intentionally corrupted data.[11] In industry, the 2023 Hollywood strikes struck back at potential corporate application. In government, policy remains under deep IP uncertainty.

At the time, many affected artists viewed this situation as potentially existential. For those at the top levels of AI policy when DALL-E first arrived, however, it was off-radar. Interviewed on the effect of image generators in late 2022, one member of the National Artificial Intelligence Research Resource Task Force, the nation's top AI policy advisory panel, had not even heard of the issue.[12] It's likely the broader task force was also in the dark.

The reason? AI had been treated as a technical specialty. Because the task force was composed almost exclusively of computer scientists, one is hardly surprised that it was not thinking about artistic rights questions. Had media policy officials recognized AI's coming general-purpose breadth, perhaps those in the arts would have been engaged in policy and had their voices heard in the design of potential solutions. A second challenge is prediction. Breadth of expertise cannot sacrifice depth of technical knowledge. Only by understanding the technical progress of generative AI—how data are scraped and used to train AI and what type of data are needed—could those concerned about artistic rights have perhaps predicted this issue and have begun to consider appropriate action. Many generators are now open sourced, meaning their code is no longer controlled by a single entity, and affected artists may have little recourse. Solving this issue would have required forward thinking, knowledgeable officials aware of technical trends, and dedicated AI policy work in media and artistic rights policy. In AI policy, such specificity is often missing.

The Importance of Deeper Understanding

To manage impactful technology, we must broadly equip officials. The National Security Commission on Artificial Intelligence recently wrote that "AI … promise[s] to be the most powerful tools in generations for expanding knowledge, increasing prosperity, and enriching the human experience."[13] All policy areas will be touched and even transformed by AI (see box 1.1). The sudden explosion in AI progress demands a new class of policymakers who not only understand AI, but also understand it in depth. Just as all policy experts need a working knowledge of economics, all will need a working understanding of AI.

Traditionally, those who *have* engaged with AI outside computer science have done so only

BOX 1.1. AI touches all federal departments

Artificial intelligence (AI) has a broad effect. One can see how it is actively affecting policy in each federal department and across disparate policy areas:

- **Agriculture:** The US Department of Agriculture is researching the use of AI to promote food safety.[a]
- **Commerce:** The Commerce Department is developing an AI risk management framework for the marketplace to provide unbiased and trustworthy AI.[b]
- **Defense:** The Department of Defense has used AI for targeting exercises and flying autonomous, unmanned aerial vehicles.[c]
- **Education:** The Education Department is seeking to engage education professionals on how AI will affect their classrooms.[d]
- **Energy:** The Department of Energy's National Laboratories researches and develops AI capabilities for many industries.[e]
- **Health and Human Services:** The Department of Health and Human Services identifies areas in the health industry that could benefit from AI, funds research to develop AI solutions, and monitors and regulates AI use in the health industry.[f]
- **Homeland Security:** The Department of Homeland Security uses AI in customs and border protection and investigations.[g]
- **Housing and Urban Development:** The Department of Housing and Urban Development is researching the use of AI risk assessments to promote fairness and equity.[h]
- **Interior:** The Department of the Interior is using AI tools to analyze wildlife, landscape, and energy information.[i]
- **Justice:** The Justice Department employs AI to analyze evidence, forecast crime, and enable rehabilitation.[j]
- **Labor:** The Department of Labor is researching the possible effects of widespread AI adoption, including the effect of AI bias on hiring and employment.[k]
- **State:** The State Department has developed and used AI to fight global disinformation.[l]
- **Treasury:** The Department of the Treasury is using AI programs to combat illicit finance operations.[m]
- **Transportation:** The Department of Transportation governs the integration of AI into automated driving systems, unmanned aircraft systems, and traffic management operations.[n]
- **Veterans Affairs:** The Department of Veterans Affairs has used AI to predict COVID-19 outcomes and reduce wait times.[o]

NOTES

a. Scott Elliott, "Artificial Intelligence Improves America's Food System," *US Department of Agriculture Blog*, July 29, 2021, https://www.usda.gov/media/blog/2020/12/10/artificial-intelligence-improves-americas-food-system.
b. Don Graves, "Remarks by U.S. Deputy Secretary of Commerce Don Graves at the Artificial Intelligence Symposium," April 27, 2022, https://www.commerce.gov/news/speeches/2022/04/remarks-us-deputy-secretary-commerce-don-graves-artificial-intelligence.
c. David Vergun, "Artificial Intelligence, Autonomy Will Play Crucial Role in Warfare, General Says," US Department of Defense, February 8, 2022, https://www.defense.gov/News/News-Stories/Article/Article/2928194/artificial-intelligence-autonomy-will-play-crucial-role-in-warfare-general-says/.
d. Office of Educational Technology, "Artificial Intelligence," accessed February 8, 2023, https://tech.ed.gov/ai/.
e. Argonne National Laboratory, "Artificial Intelligence: Accelerating Science, Driving Innovation," accessed February 9, 2023, https://www.anl.gov/ai.
f. US Department of Health and Human Services, "HHS Artificial Intelligence (AI) Strategy," December 22, 2021, https://www.hhs.gov/about/agencies/asa/ocio/ai/strategy.
g. John Hewitt Jones, "DHS Launches Public Survey on Use of AI," *FedScoop*, November 10, 2021, https://fedscoop.com/dhs-launches-public-survey-on-use-of-ai/.
h. Office of Policy Development and Research, "Using Artificial Intelligence to Promote Equity in Home Mortgage Access," *PD&R Edge*, November 9, 2021, https://www.huduser.gov/portal/pdredge/pdr-edge-featd-article-110921.html.
i. Bureau of Safety and Environmental Enforcement, "Safety Performance Enhanced by Analytical Review," accessed February 9, 2023, https://www.bsee.gov/what-we-do/offshore-regulatory-programs/safety-performance-enhanced-by-analytical-review-spear.
j. National Institute of Justice, "Artificial Intelligence: Applying AI to Criminal Justice Purposes," accessed February 8, 2023, https://nij.ojp.gov/topics/artificial-intelligence.
k. Nathan Cunningham, "How Artificial Intelligence Affects Workers with Disabilities: A New Toolkit for Businesses," *US Department of Labor Blog*, November 1, 2021, https://blog.dol.gov/2021/11/01/how-artificial-intelligence-affects-workers-with-disabilities-a-new-toolkit-for-businesses.
l. US Department of State, "Artificial Intelligence (AI)," accessed February 8, 2023, https://www.state.gov/artificial-intelligence.
m. Perkins Coie, "US Treasury Highlights Anti-Money Laundering Priorities in 2022 Illicit Finance Strategy," May 26, 2022, https://www.perkinscoie.com/en/news-insights/us-treasury-highlights-anti-money-laundering-priorities-in-2022-illicit-finance-strategy.html.
n. US Department of Transportation, "U.S. DOT Artificial Intelligence Activities," September 23, 2019, https://www.transportation.gov/AI.
o. Mike Richman, "New VA Tool Uses Artificial Intelligence to Predict COVID-19 Patient Mortality," VA Research Currents, June 28, 2021, https://www.research.va.gov/currents/0621-New-VA-tool-uses-artificial-intelligence-to-predict-COVID-19-patient-mortality.cfm.

at a basic level, a so-called Level 1 understanding. They can engage with the concept, and perhaps entertain abstract effects, but cannot dig into problems or imagine specific solutions. AI is maturing, and policymakers should go deeper. The goal should be a Level 2 understanding, in which policymakers understand conceptually how AI works and the array of core concepts and technologies on which it is built. Although they might not be able to code a neural network, they know how one functions. Although they have not studied electrical engineering, they understand the AI chip deck.

With a Level 2 understanding, this new class of policymakers can meet engineers halfway. More specifically, they will have the confidence to ask the right questions, the ability to understand engineers' explanations, and, crucially, the capability to question technical experts. This level of understanding brings AI down to earth, allowing policymakers to see the breadth of AI's effect and the many technical tools on which it is built.

How to Use This Work

The goal of this work is to equip a diversity of policymakers with the core concepts needed to acquire a degree of understanding. While a Level 2 understanding is the goal, in each section we offer two levels of depth to support readers who want only a basic understanding and those seeking greater depth. For clarity, we are identifying Level 1 material as *Fundamentals* and Level 2 material as *Deeper Dive*.

Note that AI is enabled not by one technology but rather by a diverse "constellation of technologies."[14] AI comes in many forms and uses a range of concepts and devices. To understand and solve diverse AI issues, readers must grasp the AI space. Primarily, this work seeks to explain how AI works through illustration. Along the way, it equips readers with key terms, fundamental concepts, and core technologies in a toolbox of knowledge that can be supplemented with application-specific expertise.

2. What Is AI?

Artificial intelligence (AI) is characterized by the following:

- Normatively, AI can be thought of as a goal—the goal of using human-designed systems to build something intelligent, often resembling the human mind. Descriptively, AI is commonly considered a technology, a catch-all for the many technologies and designs that make AI possible.
- AI systems generally aim to automate intellectual tasks normally performed by humans.
- Technologies such as machine learning are used to create AI systems.
- Most AI systems are best conceived as advanced inference—or prediction—engines. These inferences are used to produce analysis, inform decisions, and take automated actions.
- AI is the result of a triad of essential inputs: software (algorithms), hardware (microchips), and data.
- The core advantages of AI systems are advanced automation, analytical speed, and greater scale of action.
- While AI systems have traditionally been geared toward narrow applications, more general-purpose systems are emerging. Despite widespread attention on these general-purpose systems, the bulk of systems in use are designed for discrete, narrow-use cases.
- An algorithm is simply a logical sequence of steps needed to perform a task. In computer science, algorithms are written in code.
- Machine learning algorithms are often trained with data stored in a databank or collected in real time.

FUNDAMENTALS

Basics of AI

"A fundamental problem in artificial intelligence is that nobody really knows what intelligence is."[15]

—Shane Legg and Marcus Hutter, Google Deep Mind

There is no *one* accepted definition of AI; there are, in fact, hundreds. For policy experts, Congress thankfully simplified definitional selection by hard coding an AI definition into law through the National Artificial Intelligence Initiative Act of 2020. Legally, AI is defined as follows:

> a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations or decisions influencing real or virtual environments. Artificial intelligence systems use machine and human-based inputs to—(A) perceive real and virtual environments; (B) abstract such perceptions into models through analysis in an automated manner; and (C) use model inference to formulate options for information or action.[16]

This definition is wordy, but a few core concepts stand out.

Intelligence

First, note that the legal definition just mentioned does not explain the goal of AI technology. The reason: the goal is in the name. As observed earlier, artificial intelligence itself is a *goal* enabled by a set of ever-changing technologies (for example, machine learning). The bounds and aims of this goal are naturally murky because there is little consensus on what constitutes "intelligence." Some believe serious research should downplay mimicking intelligence, and specifically human intelligence, as the end goal, with emphasis alternatively placed on advanced task automation, data analysis, and other goals to set expectations. That said, several organizations today, such as OpenAI, are explicitly seeking to produce generally intelligent, human-level (or beyond) systems. Mimicking human intelligence was the goal of AI's founders, and most watershed moments in AI history such as AlphaGo's mastery of the game Go involve outmatching specifically human intelligence. Although defining intelligence is murky, there is no question that many AI engineers (for better or for worse) will keep some notion of human intelligence as their ultimate goal.

Readers should take this approach with a grain of salt. Focusing too intently on efforts to mimic human thought can distort our understanding of what an AI system is or represents. It also distracts from progress in the many systems that aren't trying to achieve human or general intelligence. Today, most AI systems are narrowly scoped, seeking complex task automation. Facial recognition systems, for instance, are not trying to create human intelligence; they are trying to automate human identification.

Regardless of the aim or application, modern AI systems are united by a general attempt to "automate intellectual tasks normally performed by humans,"[17] an effort naturally shaped by the application at hand and the personal views of its engineers.

Inference

A second highlight from this definition is that machine-based systems "make predictions, recommendations or decisions." In the field, this is

called **inference**. Inference is at the core of all systems, and the goal of AI systems can be generalized as the goal of making good inferences. When one asks the Alexa voice assistant to play a song, it infers a song title based on the sound of the words, triggering instructions that compare that inferred title against other titles in its database. It then plays the most likely match.[18] Similarly,

- Identifying and labeling the contents of a picture means inferring the correct match between the input image and potential labels, and
- Autonomously operating a car requires thousands of near-instant inferences about which actions to take in the near future—that is, predictions based on the position of the vehicle and surrounding objects and other information.

When these inferences trigger machine action (such as playing a song or steering a car), AI achieves the goal of automation.

The AI triad

A third highlight is the phrase "machine-based systems." AI scholar Ben Buchanan explains that "machine learning systems use computing power to execute algorithms that learn from data."[19] This is the **AI triad**:[20] algorithms, data, and microchips—the core input technologies that *together* enable AI. An essential theme of this introduction is that each of these technologies is equally necessary and interdependent. Understanding this interdependence is key to designing AI policy.

Benefits of AI

Before diving into how AI works, one must form an idea of what AI systems offer:

1. **Automation.** AI can automate new types of tasks that previously required human input. Before AI, automation was reserved for the consistent, predictable, and repetitive.[21] AI expands automation into "fuzzy" tasks that deal with complex problems and uncertainty. With AI, automation can extend to imprecise tasks, including image recognition, speech translation, and writing.
2. **Speed.** AI can resolve complex problems nearly instantly. Driverless cars face no cognitive lag when responding to hazards, and ChatGPT faces no analysis paralysis when writing. Decisive, near-instant decisions provide an advantage over human decisions, which can lag as a result of indecision, stress, and other factors. In other cases, speed can also be a hazard of its own. An extreme example lies in military systems that once granted autonomy over target engagement, allowing action before a human commander authorizes engagement.
3. **Scale.** AI can effectively perform certain tasks better than an army of humans hired for that purpose. For instance, streaming can simultaneously address the individual preferences of millions of music listeners or TV viewers, drug discovery systems can analyze millions of compounds, and ChatGPT can search and connect millions of disparate ideas.

System Flexibility

Today, all AI systems can be categorized as **artificial narrow intelligence**,[22] designed to perform a specific, limited function set.[23] These AI systems can perform one or a few tasks with high

quality but cannot perform tasks outside their discrete training.

AI applications range from single-purpose systems, such as OpenAI's DALL-E image generator, to more complex systems, such as driverless cars or even ChatGPT. Even within these narrow domains, AI can still suffer inflexibility. **Generalization** refers to a system's ability to "adapt properly to new, previously unseen data"[24]—that is, it can flexibly adapt to novel scenarios it hasn't been explicitly trained to handle. The more a system can generalize and deal with the unexpected corner cases in its domain, the higher its quality. Imagine a driverless car that is highly accurate but only in average, fair-weather road conditions. This car would perform perfectly in the majority of cases, yet when it meets a rare and unexpected situation—say, a tornado—it may not know the best course of action to protect the driver.

Although today's AI systems are narrow in scope, efforts are under way to develop so-called **artificial general intelligence** (AGI), which has "the ability to achieve a variety of goals, and carry out a variety of tasks, in a variety of different contexts and environments."[25] This category represents the science fiction vision that many readers hold of AI. Note that generality does not imply balanced quality across capabilities. Just as a lion might excel at hunting and a human at mathematical reasoning, it is possible for AGI systems to perform tasks at varying levels of proficiency.[26] Also note that AGI does not imply humanlike AI; AGI can be as advanced as humans without necessarily mimicking our cognition.[27] A chess-playing AI, for instance, might win by mere exhaustive calculation of every combination of possible moves. Contrast this thought process with the strategic reasoning of human cognition. AGI also does not mean **superintelligence**—that is, an AI system that is smarter than humans in almost every domain.[28] These variations on advanced AI systems do not yet exist, though increasing R&D has been devoted to their development.

Policymakers should take these concepts seriously even if they consider true AGI far off or impossible. Even an AI that can convincingly mimic AGI or superintelligence ought to be a matter of policy concern.

How AI Works: Prerequisites

The following sections discuss the various elements of the AI triad and the way AI works. First, several basic terms and concepts are as follows:

- **Algorithm.** "A logical sequence of steps to solve a problem or accomplish a task."[29] Although this term sounds to some like technical jargon, algorithms are everywhere. For instance, Grandma's pot roast recipe is a type of algorithm: a list of steps that, if followed, can produce the delicious Sunday dinner. In computer science, this term is more specific, referring to the list of instructions, or **code**, that a computer follows. The essence is still the same; the computer follows lines of code to perform its tasks just like one might follow a recipe. The term is often used interchangeably with **computer program** and **software**.

  Although this guide defines algorithm in its most general sense, in the context of AI, "algorithm" is often used as shorthand to refer more specifically to machine learning algorithms, the processes that a computer follows to create artificially intelligent software.
- **Model.** Unlike the more general term "algorithm," the model is the software

configuration that, once fed input data, can produce output inferences, predictions, and decisions. The model is the end result, which is the inference software created from the iterative refinement of machine learning or engineering.[30] When one trains an AI system, one is training the model; when one runs an AI system, one is running the model.

- **Machine learning.** Most AI systems today are the result of a process called machine learning. Machine learning is a method for iteratively refining the process a model uses to form inferences by feeding the model stored or real-time data. This learning process is called **training** and is a necessary step to build artificially intelligent systems. In section 6, "Algorithms," this process is explained in greater detail.

In addition to understanding what AI is and how it works, many policymakers must know how to assess it. Unfortunately, there is no one performance metric for AI models, and the measurement criteria used are highly specific to each application and are constantly changing. This study offers a starting point, describing several common metrics and the way to approach these figures with a critical eye.

DEEPER DIVE

Accuracy Assessments

A natural starting point for quality assessment is **accuracy**, which measures how a system's inferences and actions match expectations. Accuracy is broadly useful, understandable, and often sufficient. Note, however, that perfect accuracy will rarely be possible. When deploying AI applications, engineers must actively decide on an acceptable rate of failure (a choice based on their own reasoning), application requirements, and perhaps regulatory prescriptions. Alexa, for instance, answers incorrectly around 20 percent of the time.[31] In Amazon's estimation, this rate of failure is acceptable. This estimation illustrates that accuracy need not be perfect when the stakes are low.

Contrast this example with safety modules in a driverless car. In this case, many argue that, given the danger, the acceptable level of accuracy must be higher.[32] Safety still must balance practical considerations. Projections show that deploying a driverless car that is only 10 percent safer than one with human drivers could still save many lives; perhaps a seemingly high rate of failure might be acceptable if it still minimizes comparative risk.[33] Other AI benefits must also be weighed against accuracy. Conceivably, driverless cars could efficiently clear traffic in the presence of ambulances, potentially saving lives. Perhaps such a benefit would justify a lower rate of overall accuracy.

Accuracy Is Not Everything

Accuracy, although an important metric, cannot fully assess system quality in all cases. For instance, if a deadly virus appears only once in a sample of 100 patients, a disease-spotting AI coded to *always* predict a negative result would still be 99 percent accurate. Although highly accurate, this system would fail its basic purpose, and the sick would go untreated. For policymakers, a critical eye is needed to ensure that the numbers provide proper nuance.

To gain a better sense of the quality of a system, one may need additional **evaluation metrics**. It is important to emphasize that any metric used to evaluate AI carries tradeoffs. As an illus-

tration, there is often a tradeoff between measuring false positives and false negatives.[34] Choosing which to prioritize in evaluation depends on context and system goals.

Returning to the disease-detecting AI example, suppose one is doing aid work for the United States Agency for International Development. The chief concern is *treating* disease, and there is no cost to treating a healthy patient. In this case, one might prioritize minimizing false negatives so as to ensure that those with the disease get treatment. Also, one might measure quality using **recall**, a metric that states the percentage of the model's negative results that are true negatives.[35] This metric would allow one to see the likelihood of a false negative, and if that probability is low, the model is effective for our purposes.

Now imagine the reverse: suppose one is an official at the Centers for Disease Control and Prevention, and the chief concern is correctly analyzing *disease transmission*. In this scenario, perhaps one would want to minimize false positives by measuring with **precision**, a metric that evaluates how many positive results of the system are indeed positives.[36] If precision is high, then one can be certain that one is correctly identifying positive results and can better track transmission.

If one finds both false positives and false negatives undesirable, perhaps one wants a model that minimizes both. In this case, one would try to maximize the **F1 score**, assessing how well the model minimizes *both* false negatives and false positives.[37]

These example metrics are widely used to assess AI that seeks to classify data; however, that is only one aspect of evaluation, and it is not necessarily ideal for all applications. Consider how one might assess the quality of art-generation software. This task is naturally fuzzy and, in many cases, depends on the priorities or tastes of individuals; this is not something that can be easily captured in statistical metrics. A 2019 study found that for **generative adversarial networks** (GANs)—an AI model that can serve as an AI art generator—there were *at least* 29 different evaluation metrics that could be used to assess the overall quality of these systems in different contexts.[38] AI evaluation metrics, like AI itself, are meaningless without application.

Benchmarks

Although evaluation metrics can usefully describe an individual model's effectiveness, they are not suited for comparing models or tracking progress toward certain goals. As such, AI researchers have adopted a variety of **benchmarks**, common datasets paired with evaluation metrics to allow model comparison and results tracking and determine state-of-the-art performance on a specific goal or task.[39] These benchmarks are often tailored to specific tasks, goals, and complexities. For instance, ImageNet benchmarks image detection and classification,[40] while HellaSwag benchmarks a **chatbot**'s commonsense reasoning.[41]

Although useful for tracking improvements in AI systems and the state of the art, these benchmarks can be limited in their descriptive abilities. Researchers have noted that while benchmarks are often seen as describing general AI abilities, what they actually represent is more limited in scope, measuring only a system's ability at the tightly constrained benchmarking task.[42] Even if an AI system is able to accurately identify most images in ImageNet's database, that action does not necessarily mean those abilities will translate to real-time, real-world image recognition. The complexity and noise of real-

world analysis can be a far cry from the limited frame of benchmarking tests. Further, it has been noted that benchmarks often fail to test necessary characteristics such as a model's resistance to adversarial attacks, **bias**, and causal reasoning.[43] Benchmarks are constantly being replaced, supplemented, or updated as these limitations are discovered.

3. AI Policy Challenges

Before digging into the technology that makes AI possible, we must first establish what artificial intelligence (AI) policy looks like today and what issues are at stake. Currently, there is limited artificial intelligence–specific law. Only a handful of federal laws relate directly to AI, and those that do, such as the National Artificial Intelligence Initiative Act of 2020, cover basic study and coordination rather than explicit regulation.[44]

There is comparatively more AI-specific policy and executive action, though this too is in introductory stages. The National Institute of Standards and Technology's widely used AI Risk Management framework[45] provides optional processes and considerations for organizations looking to responsibly and safely develop and deploy AI systems. The 2022 Blueprint for an AI Bill of Rights lays out a list of principles officials believe should guide AI application and policy.[46] More substantially, 2023's Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence acts as the guiding document of AI strategy in the United States. This lengthy list of requirements includes limits on government use of AI, and it requires chief AI officers in most agencies, AI talent development initiatives, and technology reporting for frontier AI labs. Beyond these specific actions are introductory requests to a range of agencies to consider research, investigation, or even actions on critical infrastructure risks, civil rights, market competition, intellectual property, AI bias, and consumer protection, among other steps.

Such executive actions depend on a range of preexisting general-purpose statutes that apply to and can regulate all technologies, AI included. For instance, 2023's AI executive order used the Defense Production Act's industrial base assessment powers as the basis for its new technical reporting requirements for frontier AI labs.[47] Likewise, the Federal Communications Commission ruled that, under the Telephone Consumer Protection Act, AI-generated voices in robocalls

are illegal.[48] Other general-purpose statutes regulating domains include consumer safety, transportation, intellectual property, healthcare, national defense, the justice system, and discrimination (among others) that likely also will apply to specific AI systems in certain circumstances.

At the state and local levels, policy is varied and often more application specific. In many states, actions have been targeted at limited-scope and well-publicized AI applications and issues. These applications include "deep fakes,"[49] AI-generated election materials,[50] autonomous vehicles,[51] and AI-assisted hiring.[52] Following ChatGPT's release, some states have considered various forms of broad, comprehensive AI regulation, though no such bill has passed. In all cases, there is clearly a desire to act (perhaps regulate) and manage certain AI risks. What that may look like and how state legislation interacts with federal legislation remain to be seen.

The design of AI law and policy is and will be a complex task because of the importance and wide reach of this technology. The following sections offer a few questions that policymakers should consider when designing AI policy.

Critical Questions for Policymakers

Policymakers face many important decisions in the areas of research, development, and manufacturing; inputs and resources; quality control; externalities; and security and safety. This section discusses each in turn.

Note that related to each issue is a broader question of implementation and governance. Because AI broadly impacts society, policymakers must consider how to structure regulatory and policy governance. They should consider these high-level questions:

1. Should there be a dedicated "AI agency," or should policy and regulation be devolved to domain-specific agencies? What problems would a potential new agency solve? What would its jurisdiction be?
2. What gaps and overlaps in law and policy might hinder clear, effective policy? How can we identify those gaps and overlaps?

Research, development, and innovation

Chip development. Historically, the US government has sponsored and supported AI chip development. The recent CHIPS and Science Act illustrates the support of the semiconductor industry by policymakers of both parties.[53] This legislation follows a long history of public engagement with this sector. While the issue has enjoyed congressional support, the utility of AI industrial policy has been the subject of considerable debate, including the following questions:

1. Is there certain fundamental AI chip research that might not exist without government support?
2. Does government support and subsidization risk crowding out or privileging certain innovations and alternative designs?
3. How can policy play a role in ensuring that US industry competes with China's considerable state-led AI investments?

Computer science and algorithmic research. Algorithm and computer science research has long been intertwined with public research support and policy. Early neural networks, for instance, were first introduced by the Office of Naval Research.[54] The Defense Advanced Research Projects Agency's Grand Challenge, a

military-sponsored desert race, sought to incentivize autonomous vehicle progress through a competition and cash prize.[55] Some argue that this race supercharged autonomous vehicle breakthroughs. Although the public history of AI algorithm development is perhaps impressive, one should note that all such policies involve tradeoffs, risks, and implementation challenges. Policymakers should consider the following:

1. How can the timeliness and efficiency of public research support be improved?
2. What form of research support, including computational resources, prize challenges, or monetary grants, will best support a given goal?
3. How might public investments crowd out or supplement private funding? Will such investments distort or privilege certain research outcomes?
4. How can one best incentivize development while minimizing market distortions?
5. How can one ensure continued national competitiveness in R&D writ large?
6. How can algorithms be developed and designed to support principles, including democracy, freedom, and fairness?
7. What types of AI and applications should the public support? Should policy focus on foundational or applied research? For military research, how does one ensure that innovations are designed for dual use?

Open source. A significant portion of AI development is open source, raising questions about safety and regulation and continued innovation. Some worry that open-source, highly capable models will leave potentially harmful software uncontrolled and easily accessible to bad actors. Conversely, some worry that attempts to regulate open source will cost innovation by depressing highly dynamic innovation of the open-source community and weaken the security benefits that transparent, easily critiquable software enjoys. Enforceability is another challenge: how can regulations apply to anonymous actors? Therefore, policymakers should consider the following questions:

1. What are the costs and benefits of open source AI? Do the cybersecurity and innovation benefits of open-source models outweigh potential risks of open models? What risks do open models pose?
2. How could any potential regulations be enforced? What are the limits of success?
3. How can governments improve open-source code? Is there a place for public open-source analysis or vulnerability tracking?
4. How can public-sector code be open sourced to share potential innovations with other governments, agencies, and actors?

Inputs and resources

Supply chain robustness. AI chips and hardware require a diverse range of materials and components to support processing needs. A robust AI ecosystem requires supply chains that can reliably source and provision the resources needed by the AI economy. Toward these ends, policymakers should consider the following:

1. How can the United States trade openly with new markets to ensure access to these goods?
2. How can the United States ensure an efficient and balanced supply chain?

3. Can domestic resources help supply the needed materials? How can the United States balance the benefits of domestic resource extraction with environmental costs?
4. How can the United States ensure access to key resources such as rare minerals? Can alternative materials be developed or discovered to reduce dependence on rare or environmentally harmful materials?

Talent and immigration. AI development requires a range of highly technical and specialized skills. Supporting manufacturing, research, design, and deployment will require a deep talent pool and expensive labor. Education, grants, apprenticeships, and immigration can help fill this gulf. Policymakers should consider the following:

1. What education policies can incentivize AI and computer science education? What type of skill sets are needed?
2. How can nontechnical fields be upskilled with AI knowledge to prepare those fields for the potential effect of AI?
3. Can private-sector incentives for training and apprenticeship programs reduce educational burdens?
4. How can immigration policy be reformed to attract and retain global talent?
5. How can AI education balance technical skills with a need for free and creative thinking?

Data resources, privacy, and intellectual property. The scale and source of data often used to train AI systems has prompted a diversity of concerns over data rights, privacy, and intellectual property. Important policy questions include the following:

1. How can the United States ensure that governments and companies adequately protect the vast and sensitive data used to create their AI systems?
2. How can the United States mitigate concerns that it will lose an "AI race" to China because authoritarian tools allow for more extensive and detailed data collection?
3. There is concern that new market entrants with limited data stores and scraping capabilities cannot compete against the vast stores of user data amassed by big tech firms. How can the United States ensure a level playing field and a competitive market?
4. How can copyright and IP law balance fair use, artistic autonomy, intellectual property rights, and continued AI innovation?

Data standards and interoperability. Data standards can affect the nature and usability of data. Healthcare AI, for instance, has been slow to develop because of highly siloed data, disparate technology practices, and recordkeeping differences across systems.[56] A key to this problem is interoperability and standardization. If technology can easily communicate and share data, and if data are standardized and easy to use, this could aid the development of AI systems. Toward these ends, policymakers should consider the following:

1. How should the government design and format data standards to best serve AI? What information should the data capture? How do these decisions affect the ability to share data, develop AI systems, and promote innovation? Conversely, how might standardization hinder innovation?

2. How can the government reduce data balkanization to ensure that AI has the tools it needs to grow? How might this be balanced with privacy and security concerns?
3. How can industry or private actors set and manage standards and data interoperability without government involvement?

Quality control

Explainability. Because AI systems often focus on capabilities rather than explanation, the reasoning behind their actions can be opaque. Law and policy often require clear reasoning and decision-making. This requirement can raise questions and concerns, such as the following:

1. Should the government risk using autonomous weapons if we do not understand how they select, and possibly kill, targets?
2. Should the government use AI sentencing algorithms if we do not know if their final decisions are affected by racial biases?
3. How does the government know that an AI's decision-making process has not been compromised by a malicious actor?
4. How might the government know if autonomous vehicles are making safe decisions?
5. How does the government know that statistical AI models are producing high-quality predictions and results?

Overfitting and underfitting. Overfitting is the problem of fitting a prediction algorithm too tightly to training data, so much so that it underperforms with new data. Underfitting, in turn, is the failure to adequately fit an algorithm to the training data, rendering predictions with new data altogether unreliable. For policy-sensitive applications, AI models must be able to demonstrate that they are neither over- nor underfit for the task at hand. At present, there is no easy solution to this challenge. For policymakers, the best approach is vigilance. The following are examples of issues that this challenge could create:

1. Economic data have a relatively short history. Treasury models therefore run an underfitting risk that could lead to faulty algorithms when trying to predict inflation, employment, and other key metrics.
2. Court-sentencing algorithms can run the risk of overfitting. If a case used in a model's training set is sufficiently unique, the model could carve out a prefabricated decision path that is not generalized but instead is tailored specifically to that set. Should this sort of model, regardless of fitness, ever be used by courts?

Bias and auditing. The data used and the bias embedded in AI algorithms can lead to incorrect or harmful results. AI-powered pulse oximeters, for instance, have been found significantly more inaccurate for dark-skinned patients.[57] Such biases can cause harm. In another case, Amazon found unintentional bias embedded in its hiring algorithm, which favored male applicants far more than female ones.[58] Such biases can cause discrimination. One proposed path forward would be AI audits that could be used to assess algorithmic weaknesses, security, and bias. Regarding bias and auditing, policymakers should consider the following questions to address these issues:

1. What algorithmic design best practices and industry standards can help spot and mitigate bias?
2. What data sourcing, cleaning, and processing standards can help minimize bias and

ensure robust algorithms? What tradeoffs, unintended consequences, or concerns could such standards create?

3. Is there an acceptable level of bias? What biases are unacceptable? How does the law deal with AI bias?
4. Can intentional bias be used to mitigate negative biases? What risks or unintended consequences could this pose?
5. Should AI audits be required? If so, when, and what processes should they include to ensure strong results? Further, would requiring audits place an undue burden on innovation?

Externalities

Energy use, emissions, and environmental impact. Supporting AI requires significant energy use. Chip fabrication requires extensive energy resources,[59] as does the computer-intensive training process. Energy requirements expand as AI algorithms and market demand grow. As a result, intensive computing can leave a high carbon footprint. Cloud computing centers also constrain local energy supplies, potentially increasing local energy prices to support often nonlocal demand. Finally, fabrication produces wastewater and toxic by-products, while cloud computing centers burn through difficult-to-recycle semiconductors. Policymakers should consider the following:

1. How can the government and private actors balance the energy use and emissions costs of AI systems against the benefits of AI innovation?
2. Can AI system innovation in energy management and climate research be used to help reduce costs and fight climate change?
3. What waste and recycling standards and policies can ensure that waste is properly managed?

Labor disruptions. Advances in AI automation may disrupt the workforce and displace certain professions. For instance, in the United States, there are more than three million truckers, a generally low-education profession that could be eliminated or transformed by driverless vehicles.[60] Other industries may feel similar strains. Although there is no guarantee that AI will lead to fewer jobs, some people will likely have to find new employment or find that AI changes the nature of their work. As such, policymakers should consider the following:

1. How can education policy be used to upskill or reskill displaced or underskilled workers?
2. How can policy ease workforce transitions and ensure that older workers are not left behind?
3. How can agencies update or remove regulations that might entrench certain labor classes despite AI automation improvements?
4. Should the government or private actors ensure redundant human skills in fields automated by AI? If so, how?

Security and safety

Cybersecurity. The introduction of AI naturally comes with a transformation of the cyberthreat landscape. New threats can be found in AI. The massive scope of certain models can make it difficult to spot vulnerabilities or bad actors. Further, data can act as a new attack surface. Data poisoning attacks seek to inject vulnerabilities

into a system through bad data or use data inputs to cause a trained system to malfunction. AI will also be used as a tool of cybersecurity. Offensively, it could be used to hunt and exploit vulnerabilities without human involvement or generate highly convincing spear-phishing emails. Defensively, AI can be used to detect intrusions, spot bugs, and stop bad actors. Policymakers should consider the following:

1. What processes can be used to detect vulnerabilities not only in algorithms but also in the data and processors that drive these systems?
2. What standards and best practices can be passed to the private sector to mitigate and minimize AI cyber risks?
3. How can the government detect and alert the public to systemic AI cyberattacks and risks?
4. How can the government encourage effective prosocial cybersecurity research and hacking?
5. How can the government ensure critical infrastructure remains secure and operational?

Supply chain security. The supply chain that supports AI technologies is long, complex, and brittle. Chips are often manufactured abroad, leaving them vulnerable to foreign influence. Data are often collected, sold, and reused. This creates novel threats and attack surfaces. Policymakers should consider the following:

1. How can the government or private actors gather intelligence about supply chain–based vulnerabilities and threats?
2. How can the government or private actors detect compromised or counterfeit chips?
3. How does the government hedge against security threats to its supply chain, such as China's threat to Taiwan—its primary semiconductor trading partner?
4. How does the government or private actors balance the need for plentiful resources with the need to minimize the influence of bad actors?
5. How can the government collaboratively work with its allies to ensure access to safe components?

Content regulation, identification, and moderation. As generative AI grows in quality, and formats such as generated video and audio mature, political scrutiny has grown. When generated content is found obscene, objectionable, or illegal, the content itself is viewed as the problem. In other cases, content use is the challenge. Already AI has been used to generate propaganda, advertisements, and misinformation. Finally, some worry that if AI-generated content isn't readily identifiable, consumers can be misled. Without the availability of identification tools or procedures, AI-based scams, deep fakes, generated misinformation, and other challenges could easily cause harm. Policymakers might consider the following:

1. What generated content might be off limits? How might any content restrictions overlap with existing law, such as Section 230 of the Communications Decency Act and the First Amendment? How might limits be enforced effectively without harming innovation?
2. Who is liable for harm related to generated media? What impact would liability questions have on safety, innovation, and deployment?

3. Should the United States restrict the use of AI-generated materials in certain media, such as advertisements or election materials? If so, how?
4. How can policymakers respond to generated spam, disinformation, or scams? How can consumers identify AI-generated media? What technologies, rules, or norms are needed to ensure that consumers and governments understand what is generated? Should generated media identification or certain authentication procedures be required?

Lethal autonomous weapons systems. AI algorithms make a reality of robotic weaponry that can select and engage targets without humans in the loop. This is no longer science fiction; such systems are already in use on the battlefield.[61] Policymakers must actively engage in the many now-practical ethical and legal implications of these systems. Questions that policymakers must answer include the following:

1. How do autonomous weapons conform to international law and the laws of war?
2. How might arms control law apply to autonomous weapons, and how might the government technically verify a potential arms control agreement?
3. What role do humans play in controlling or mitigating the potential harms of autonomous weapons?
4. How can the actions and life-or-death decisions made by autonomous weapons be justified or explained?

Incomplete and Ever-Evolving List

This list is not comprehensive but rather a small selection of the issues at stake. The hope is that this starting point can help readers understand the importance of AI technology and its relationship to a broad array of policy domains. As they dig into the technology that makes AI possible, readers are encouraged to imagine further unanswered questions and connect these concepts to issues in their given fields.

4. Data

Data serve two high-level purposes in artificial intelligence (AI) systems. First is the input. Data are the digital raw material used to train **models** during the machine learning process, as well as the input on which trained models make inferences. Second is the output of **inference** that serves the practical purposes of users and output that can be recycled as input for further refining model performance.[62]

Several design choices of the dataset—such as volume, data selection, and the removal of outliers—shape the nature of AI systems. The technical form of digital data files also matters. The resolution of a photo, the compression of digital music, and unseen metadata all shape what information an AI system can process during learning or inference. To understand how microchips and **algorithms** shape AI, policymakers must first grasp the fundamental importance of data.

Data have many important aspects:

- Through the training process, machine learning models use data to refine their inferences.
- When deployed, trained models use input data to make inferences, which can be translated into predictions and decisions.
- Many machine learning approaches require large volumes of data to train AI models.
- Machine learning approaches with small data are emerging to enable success without big data.
- The variety of data can be just as important as the volume. With diverse and representative data, systems can better account for real-world diversity and complexity.
- A diversity of data storage, warehousing, and collection systems is an important consideration in understanding AI governance.
- The data used to train and operate systems are often the result of human curation,

labeling, and cleaning. Human curation of data can lead systems to reflect biases. Some systems may perpetuate negative biases, whereas some might be more objective.

FUNDAMENTALS

Data Volume

"We don't have better algorithms. We have more data."[63]

—Peter Norvig, director of research at Google

Whether an AI system is in development or in use, the quality of its data is paramount to success. Selecting high-quality AI data is challenging, a function of multiple competing factors including volume, variety, and velocity. These qualities together are sometimes referred to as the "**three V's**."[64]

Determining the ideal data *volume*, or the quantity of data relative to the model's needs, has become a central question of machine learning. To train AI systems, there are two emerging approaches: **big data** and **small data.**

The big data approach is likely the most familiar. To train an AI system, vast stores of data are funneled into the model, which learns from that data and refines itself over time. Although this process does not always work in practice, the hope is that with enough data, the model eventually arrives at an optimal form with powerful predictive capabilities.

The famed ImageNet database illustrates the power that large and diverse datasets can provide. Introduced in 2009, ImageNet included more than 14 million images and was conceived on the premise that progress in AI image recognition was a matter of more data, not improved algorithmic design.[65] This approach proved successful. Massive data accelerated the improvements in computer image recognition; the accuracy of models using ImageNet jumped from a modest 72 percent success rate in 2010 to 96 percent in 2015, an accuracy rate exceeding average human success achieved in just five years.[66] Such results are rooted in the volume of this database.

Although the ImageNet approach to image recognition benefited from millions of data points, the exact volume required for machine learning training is not standardized. Note that image recognition is a narrow, single-purpose application of this technology, yet it still required vast troves of data. For more complex systems, such as driverless vehicles or chatbots, the volume of data is likely orders of magnitude larger. Estimating how much data are enough is a moving target and heavily depends on the application complexity,[67] model size,[68] accuracy requirements, and other goals. Progress has been made toward defining the relationship between algorithms and data requirements;[69] however, current models are still speculative.[70] In practice, engineers often depend on soft rules of thumb rather than empirically tested processes.[71] Today's AI engineering is more an art than a science.

Trending against big data approaches are the increasingly common **small data** strategies.[72] These can be used in scenarios where data are limited, spotty, or even unavailable. Small data strategies use a variety of techniques to overcome data limitations, including **transfer learning**, where a model "inherits" learned information from previously trained models; **synthetic data**, where representative yet generated data are synthetically created;[73] and **Bayesian methods**, where models are coded with "prior information" that provides problem context before learning begins, thereby shrinking the overall learning

challenge.[74] Given predictions that high-quality data may soon be exhausted,[75] such strategies could augment or maximize the value of existing human-generated data.

In 2018, DeepMind's AlphaZero demonstrated how an AI system could master chess, Shogi, and Go through self-play—learning without *any* input data apart from the game rules.[76] The system bested all existing big data–trained systems, challenging the assumption that more data are always better. Although AlphaZero's design is not universally applicable, it demonstrates the potential of small data AI to transform future AI development. This ability to make accurate inferences without having seen explicit examples is called **zero-shot learning**,[77] while **few-shot learning**[78] is the ability to make inferences based on a handful of examples. Both are considered rare yet highly desirable qualities essential for model flexibility and robustness.

Data Variety

Variety is just as important as volume. The problems that AI systems face are often complex and, in theory, a great variety of data can help models account for the unique wrinkles and corner cases that complexity brings. Flexibility is essential to AI quality and ensures that systems are robust in the face of the unexpected. A classic example illustrating the importance of variety is the facial images used to train facial recognition algorithms. Human faces come in many varieties, and to perform accurately, an algorithm should be trained on data containing a full variety of races, genders, hair colors, and so forth. Without full variety, these systems have been shown to misidentify nonwhite faces at significantly higher rates.[79] Insufficient variety can create performance-degrading bias.

The variety of data must match the task at hand. The maps, visual images, and proximity sensor data needed to train a driverless car will be vastly different from the data required to train a stock-trading AI. Data must also be timely. Adding **stale data**—that is, old data that are not quite pertinent to the current problem—just for the sake of greater volume can reduce the overall quality of an AI system.[80] As an illustration, inflation data taken before 1971, when the US government promised a fixed rate for gold coins (and gold bullion), may contain more noise than signal for inflation data since 1971. Perhaps such data should be excluded when training economic modeling systems.

Data Velocity

Velocity refers to the speed "in which data is generated, distributed, and collected."[81] In general, this speaks to an AI system's ability to manage and access the data it needs for optimal performance.

Data generation and collection depend on the design of a system and the way it interfaces with the world. Web applications are well known for their ability to amass diverse and incisive data from their users. Meta and Google use digital platforms, social media, and adware to track users and collect personal data. Mass data collection is also widespread outside of the internet. In healthcare, electronic health records have enabled the collection, digitization, and aggregation of bulky tranches of data. These data include physician documentation, patient inputs, external medical facilities transmissions, and direct transmissions of medical data from hospital instruments. As in social media, aggregated healthcare data can be truly massive.[82]

As AI models are embedded into physical systems such as cars and drones, an "AI system" has broadened to include the visual, audio, and signal arrays that capture real-time information to function adequately. Some refer to this as the broader "AI constellation."[83] AI increasingly takes advantage of the **internet of things** (IoT), a network that connects uniquely identifiable "things" to the internet, where the "things" are devices that can sense and interact according to their hardware and software capabilities. IoT devices can prove rich data sources and give AI additional eyes and ears into a problem. Recall that one of the primary benefits of AI is its sensory scale and scope. The IoT is a relatively new phenomenon, and these devices may grow in importance to AI because they are able to collect a wide variety of previously inaccessible data.[84]

Web storage and networking technologies are essential components of many AI systems, physically distributing the data storage and processing burden. These devices include not only **data warehouses**—large, centralized warehouses holding hundreds of servers on which vast lakes of data are stored[85]—but also smaller caches of data physically closer to where the program is running to allow for quick data access. For an AI to learn quickly and function efficiently during inference, data must be easy to collect, store, and access.[86] Distributed data resources allow systems to take advantage of storage and processing power otherwise unavailable, while resource consolidation in warehouses can lead to economies of scale and lower cost burdens.

Data Management

Data management often dominates AI design. In fact, engineers frequently cite that data preprocessing accounts for 80 percent of engineering time.[87] Data are often disjointed, messy, and incomplete. Before a model can be trained, **data cleaning**, often by hand, is required to ensure usability.[88] To prepare data, engineers must decide whether to remove outliers, and they must weed out irrelevant information and ensure that the data are well organized and machine readable. Various methods and rules of thumb have also been developed to help fill in data gaps as needed.[89] To reiterate, AI engineering is often more art than science. Further, data must often be labeled. AI cannot naturally know the labels and symbols that humans apply to objects. An image of a red, shiny fruit can be labeled "apple" only if AI knows that term. All these labels are often affixed by hand.[90] Automatic label generation, however, is increasingly common. OpenAI's DALL-E 3 used 95 percent AI-generated captions in its training dataset, often enabling longer, more descriptive captions than those written by humans.[91] While traditionally time consuming and human driven, data management overall is increasingly automated.

Bias

A common concern in AI is **bias**, defined generally as the difference between desired outcomes and measured outcomes. Data are a major source of AI bias. When a model learns from human-curated data, the model takes on a lens that reflects the viewpoint of the humans who selected and shaped those data. For instance, natural language–processing AI trained on news articles may take on and perpetuate the societal stereotypes embedded in the language and viewpoint of those articles.[92] Even after training, the data input used during model inference can bias its output. Input data that ask an AI chatbot to

write a "positive poem," versus just a poem, will bias results in a positive direction.

The National Institute for Standards and Technology (NIST) notes that AI bias has many roots. In many cases, bias simply stems from the natural blind spots in human cognition and judgment and the consequent choices that engineers make about what data are more important or less important.[93] Because humans collect data, all data will be biased in some way. In other cases, bias is rooted in structural constraints. Perhaps the dataset an engineer uses is selected not because of its superior quality but merely because it was easy or cheap to access. The resulting AI system will then take on the qualities of that set, whatever they may be.[94] Historical data trends can also bias present-data AI systems. For instance, if an algorithm used to judge recidivism was trained on data marred by historical racism, its decisions could incorporate those historical prejudices moving forward.[95] Beyond these examples, there are many additional sources of bias, all of which must be balanced when selecting data.

Bias, although unavoidable, is not necessarily harmful. Often, the intensity of a given bias may be negligible or irrelevant to the goals of a system. A chatbot that is biased toward using an overly academic tone might be useful as a research reference tool despite occasionally sounding pompous. In other cases, biases may exist yet have little effect on system performance because they are exceedingly rare. Identity-based biases may be considered negligible in a system if they occurred only once every trillion queries. In all cases, engineers decide, consciously or not, what constitutes an acceptable level of bias before they deploy these systems. Today, these decisions are increasingly shaped by various AI bias correctives. Developing these fixes is inherently challenging, however, and has become a prominent focus of recent AI research and policy discussion.[96]

Beneath the stored information lies a wealth of technical decisions that decide what information is contained in the dataset, how it is to be used, and how it interacts with the AI model. A deeper understanding of the choices behind data design can reveal the lens through which AI "sees" the world. These choices can matter for policy, not only because many data standards are mandated by law, but also because they can influence or even dramatically change outcomes. The following sections introduce several concepts pertinent to the governance of data.

DEEPER DIVE

Adversarial Machine Learning

Data affect not only AI system design but also system security. **Adversarial machine learning** refers generally to the study and design of machine learning cyberattacks and defenses. Of central importance to many attacks are data.[97] The design of these systems is often driven by training data, and training data alterations made by malicious actors have been demonstrated to both degrade model performance and purposefully misdirect it. So-called **data poisoning attacks** can be implemented in some cases with only minor alterations to data. One study found that a single altered image in a training dataset caused a classification system to misclassify thousands of images.[98] As a result, poisoned data can be difficult to spot, lowering the bar for attacks.

Data can also be used to attack systems after training is complete. For instance, **adversarial examples**, which are data inputs designed to trick AI systems during inference, can cause

models to make incorrect predictions.[99] In a classic example, the addition of only a few stickers to a stop sign caused a visual classification system to classify it as a 45-mile-per-hour sign.[100] Similar attacks have been developed for a range of other AI applications.

Adversarial techniques can be used defensively. Glaze is a system that imperceptibly alters digital art, causing image-generation models training on that data to misinterpret the data's contents and style. This system is used by artists to block image generators from learning and copying their style.[101] DeFake is a system that alters human voice recordings to disrupt bad actors trying to clone someone's voice to carry out synthetic voice fraud.[102]

Beyond these prominent examples, there are many emerging adversarial attacks and defensive-use cases, and this new field of study is constantly changing. Mitigating and preventing these vulnerabilities will prove a major challenge as AI capabilities improve and become even more widespread.

Data Standards and Data Capture

Much of data that are collected and used are constrained or guided by **data standards** set by industry or government.[103] For instance, accounting data standards in the United States are set by the Financial Accounting Standards Board, which dictates how financial statements are structured and recorded.[104] Standards can deeply shape what data are available for any particular AI application. Under the board's rules, companies can pick one of three methods to account for inventory, whereas entities regulated under International Financial Reporting Standards have only two permitted methods.[105] As a result of these policy choices, the inventory data that are recorded can vary substantially.[106] If applied to AI, these data differences can ultimately alter analysis and results. As with all concepts in AI, application matters. The effect of some standards may be minor in certain cases but dramatic in others.

Standards dictate not only the content of data but also the structure of their digital **representation**. MP3, PDF, and other **file formats** are familiar to most people. Each of these file formats is a standard that dictates how to arrange 1s and 0s to properly represent a given piece of data—in the case of PDF, a document, or in the case of MP3, an audio file. These formats can affect the quality of data and, by extension, AI. For instance, some formats, such as JPEG, allow for image compression, a technique that seeks to reduce file size by removing data from an image. This approach can have significant implications. In mammography image analysis, results have been found to vary significantly when AI systems are trained on images of differing compression levels. In certain cases, compression even caused complete misinterpretation of mammograms.[107] It is worth repeating: data standards are design choices that are critical to AI applications.

Furthermore, note that, increasingly, captured data are not necessarily free from AI influence. Many cameras, including the cameras in the Apple iPhone, employ AI techniques to subtly alter images during capture.[108] Although the effect of these alterations remains to be seen, what is captured in data does not necessarily represent the unaltered ground truth of reality.

5. Microchips

In 1997, the addition of a tailormade "chess chip" allowed IBM's Deep Blue artificial intelligence (AI) system to defeat world champion Gary Kasparov in chess.[109] This defining moment in AI history was enabled by improvements in the engineering of semiconductors and the manufacture of microchips (or simply, chips). Since then, a recurring theme in AI innovation has been the importance of ever more efficient chips. Without the significant improvements in microchip capabilities since 1997, none of the **big data** or **machine learning** strategies that have supplanted the more primitive AI methods used by Deep Blue would have been possible.

Microchips serve two primary purposes in AI: providing processing power and storing data. Perhaps their most important quality, however, is the speed that enables quick computation and, by extension, intelligence. This section discusses how microchips function and addresses the increasing importance of this element to AI innovation.

Microchips have many important aspects:

- AI systems depend on microchips to run AI algorithms and store data.
- Variations in chip design can offer unique functions, speeds, and storage properties to AI systems.
- Chips are increasingly AI specific. Popular AI-specific designs include graphics processing units and application-specific integrated circuits.
- Over the past four decades, microchips have improved exceedingly quickly, doubling their processing speed roughly every two years by physically shrinking computational units. Owing to physical limits, however, this geometrical pace may not be sustainable over time. Future chip innovation will depend on architectural innovation.

- Microchip design and manufacturing are complex and supported by a wide range of disciplines, technologies, and companies.

FUNDAMENTALS

Microchip Basics

Although separate concepts, microchips are often refered to as **semiconductors**. The name "semiconductor" comes from **semiconductor materials**, such as silicon or germanium, the key ingredient in chips.[110] Chips contain many components, but their power and speed are owed to their **transistors**, the semiconductor switching device that performs computation. As a rule, chip power and speed increase as the transistors on a chip shrink in size and grow in density—that is, more transistors fitted into the same space. Historically, chip innovation has been linked to transistor innovation (specifically, transistor size reductions). For decades, consistent transistor improvements have unleashed the ever-growing processing speeds that, in the 1990s, enabled systems such as Deep Blue and, in the modern era, machine learning.

Chip innovation has long followed a pattern, known as **Moore's law**, in which the number of transistors per chip doubles roughly every two years.[111] Less a "law" and more an observation of chip innovation patterns, Moore's law has nonetheless held true over recent decades. The resulting pace of chip improvement has allowed for predictable improvements in the design of AI systems. For **algorithms**, this improvement has enabled greater processing speeds and therefore quicker "AI thinking."[112] For data, this advancement has built the storage capacity needed to support big data.[113] Transistors, however, are shrinking to their physical limits, and their performance no longer will advance as quickly, if at all, on the basis of size alone. Future improvements in chip function, and by extension AI, will require innovation beyond shrinking the transistors inside microchips.[114]

AI Chips

The past stability in the rate of growth of processing power meant that AI research focused on algorithms, sidelining discussion of hardware. In recent years, however, hardware has been at the center of the AI conversation. To meet processing demands, researchers are turning to **AI chips** (also called **AI accelerators**), a range of chips that are designed specifically for the unique processing needs of AI.[115] AI chips improve performance not through transistor size reductions but via changes in **microchip architecture**—the "blueprint" configuration of chip components.

The AI chip advantage is rooted in speed and specialization. **Central processing units** (CPUs), the general-purpose chip used for AI before the emergence of AI chips, are flexible but less efficient than AI-dedicated chips when processing AI-specific calculations.[116] CPUs perform inefficiently when operations are repeated in bulk and when memory is frequently accessed, which are requirements of most AI algorithms.[117] AI chips can solve these problems.

In brief, in addition to CPUs, there are currently three categories of AI chips that policymakers should understand: **graphics processing units** (GPUs), **field-programmable gate arrays** (FPGAs), and **application-specific integrated circuits** (ASICs). GPUs, FPGAs, and ASICs can be conceived of as standing on a spectrum spanning greater flexibility at the GPU end and greater speed at the ASIC end, with FPGAs standing in the middle (figure 5.1).[118]

FIGURE 5.1. The semiconductor speed-flexibility tradeoff

GPUs are limited-purpose chips originally designed for graphics processing, but they have been appropriated for AI.[119] Training a neural network, the most common AI model, requires large-scale and frequent matrix multiplication, a simple yet time-consuming mathematical operation.[120] GPU architecture is designed with many matrix multiplication units that can execute multiple operations simultaneously, a quality known as **parallelism**.[121] To analogize, a CPU is like an expert chef—versatile at cooking any dish simple or complex, though limited in that he or she can cook only a handful of dishes a night. GPUs are like an army of fast-food cooks—their versatility is low, and the food is not complex, but through raw numbers and focusing on only a few menu items, these cooks are able to feed far more people each night.

FPGAs and ASICs are single-purpose chips custom built for each application. In both, the AI software is hard-coded directly into the chip's silicon base. Application specificity increases speed by removing unneeded features and streamlining computation. The core difference between the two is programmability: the circuits baked into FPGAs are custom built *and* can be updated as needed. Meanwhile, ASICs are custom built but cannot be updated.[122] FPGAs, owing to their programmability, carry certain efficiency costs. ASICs are perfectly tailored to an application's specific needs, giving them greater speed.[123]

Chip selection depends both on the phase of AI deployment and on application-specific inference demands. GPUs, owing to their flexibility and parallelism, command the vast bulk of chips used to train systems.[124] During inference, however, application-specific demands have led to greater diversity. For speed-critical applications, such as real-time monitoring systems, the superior speed of ASICs can be critical. For some consumer products, pricing is key, favoring CPUs over comparatively expensive AI chips. Overall, the growing trend in AI inference chips is a steady gain of market share by ASICs.[125]

Often, **floating point operations per second** (FLOPS) are used to measure processing power and intensity. A floating point operation is any basic mathematical operation (addition, multiplication, etc.) on rational numbers (that is, numbers with decimal points; for instance, 3.12). Measuring the number of operations per second gauges how quickly processors run computations and programs. Confusingly, floating point operations—FLOPs, with a lowercase "s"—are also used to measure model "size" based on how many operations that model requires. Both are common in policy and computer science literature.

DEEPER DIVE

Microchips in Detail

What makes silicon and the other semiconductor materials that power computing unique is their

ability to act as both insulators and conductors, depending on certain conditions.[126] This quality is significant because it allows engineers to program exactly *when* these materials will conduct electricity. The working part of chips made of semiconductor material is the transistor. Functionally, a transistor is an electronic switch that alternates from allowing current to flow to blocking current. When current flows, it is represented as a 1, and when it is blocked, it is represented as a 0. This core function forms the basis of data representation and computation.

Transistors are built from a combination of silicon and **dopants**, impurities that alter the properties of conductivity to enable engineers' discrete control over electric currents.[127] Without dopants, engineers could not control when and why a transistor switches on or off.

To manipulate and store electrical currents, one can link transistors together in **circuits** that enable them to perform basic computation. For instance, an adder is a common circuit that takes in two numbers and adds them together. Transistor circuits can also form **memory units**. For instance, static random-access memory (SRAM), a type of computer memory, uses a small collection of linked transistors to trap energy, thereby storing the data that energy represents.[128]

Integrated circuits (ICs) are devices that string together many of these circuits, memory units, and other peripheral components to create a toolbox of basic operations that software engineers can use when running algorithms. ICs often include **execution units**, subsystems that package related circuits together with memory and other tools to enable basic functions. These execution units come in many forms, each with a specifically designed purpose. An arithmetic logic unit, for instance, may include an "adder" to perform addition, as well as all other circuits required for basic arithmetic.[129] The toolset provided in a chip can vary widely, and supporting AI often means choosing chips with the ideal set of capabilities.

Chip Design and Manufacturing

Central to many policy questions are issues related to the design, manufacture, and supply chain of microchips. These systems are highly complex, and they are supported by a wide web of technologies and engineering disciplines. Ensuring AI innovation naturally involves ensuring a robust and secure supply chain.

Talent

The skills required to develop AI chips are fundamentally different from those needed for AI algorithms and data management. The scientists who design AI chips tend to be electrical engineers by trade; algorithms and data are the specialty of software engineers.[130] Further, manufacturing requires an even more distinct skillset to develop the physical processes, machines, and production foundries. This requirement expands the necessary AI talent pool to include an array of disciplines, including chemical engineering, materials science, and mechanical engineering.[131] AI innovation is not the domain of computer science alone.

Development and fabrication

Microchip development goes through several core phases. To design a chip, engineers wield **electronic design automation** (EDA) software that allows them to map a chip's execution units and arrange transistors.[132]

Once designed, chips are then fabricated in foundries where chips are not assembled

but printed. In brief, the process starts with a **wafer**, a raw chip base, usually made of silicon. Next, a variety of materials are printed onto the chip to enable **photolithography**, a process by which light is shined through a "circuit stencil" known as a photomask, printing the design onto the chip. Additional elements are added through **etching** (using chemicals to remove unwanted material and shape the design) and **deposition** (blanketing the chip with materials to add components).[133] The list of materials required spans a large portion of the periodic table. Therefore, manufacturing requires an extensive supply chain, materials stock, and a chemistry knowledge base to support manufacturing operations.[134] After chips are printed, they are packaged in a protective casing and shipped.

Material science innovations are an often overlooked source of greater AI processing power. For instance, engineers have found that using thinner ultraviolet rays, rather than visible light, in photolithography can embed chips with thinner components, decreasing chip size and increasing chip speed.[135] To reiterate, AI innovation is not the domain of only computer science.

As a generality, the equipment used in chip development and manufacturing is highly specialized and, as a result, highly expensive. Photolithography scanners, for instance, can cost more than $100 million per unit.[136] Specialization has also led to concentration. In some cases, this concentration is geographical; for example, as of 2022, 85 percent of leading-edge chips were manufactured in Taiwan and the remaining 15 percent in South Korea.[137] The Dutch firm ASML Holdings is the only manufacturer of the extreme UV lithography machines needed to make all state-of-the-art chips in use today.[138] All of these factors complicate the robustness and security of the AI supply chain and have recently received significant policy scrutiny.[139]

Hardware infrastructure

Once these chips are produced, their specific arrangement and use in AI systems are also essential to the power they unleash. Not all these hardware capabilities will be housed locally. **Cloud computing**, a general concept in which computing resources are stored remotely and can be accessed for a fee, helps provision resources. The cloud cheapens computational cost through economies of scale and lowers the barrier to entry for AI. This approach can allow researchers to access the resources they need without buying physical semiconductors.[140] Naturally, this framework renders both the AI supply chain and the AI regulatory puzzle ever more complex. Pieces of an AI system can exist in multiple locations that collectively provide needed resources. Decentralized computing techniques such as **federated learning** further muddy the waters by eliminating centralized computing and data storage. This technique trains AI systems on a web of disconnected servers, rather than a centralized server, to eliminate data aggregation and preserve privacy.[141] Such techniques could add regulatory complexity by eliminating the ownership link between AI engineers and the data they use.

6. Algorithms

Artificial intelligence (AI) algorithms serve two main functions: inference and learning. The goal of AI **models** is to produce statistical inference based on data—data for training the model and new data. For instance, a chess-playing AI system must infer the chess move that, from all available moves, is most likely to lead to victory. During learning, models improve their performance through iterative data analysis, which is known as training or, more narrowly, machine learning.

This section introduces algorithms. It discusses varieties of **models**, the way they learn, the way they perform inference, and the key challenges inherent in their application and design.

Key points:

- Most AI algorithms are varieties of machine learning, a technique that produces intelligent systems through learning from input data or direct experience.
- There are several variations and approaches to machine learning.
- Neural networks are perhaps the most common technique used in designing AI models, including current cutting-edge applications.
- As with the choice of data, the choice of algorithmic technology can both influence and bias results.
- Many AI systems are opaque, and the process that leads to their predictions and decisions is often difficult to explain. AI explainability efforts are under way to render these processes transparent and understandable.
- To promote AI quality and safety, many propose AI audits that would assess the biases, accuracies, and strengths of systems before and while they are deployed.

FUNDAMENTALS

Varieties of Machine Learning

Machine learning is a method for iteratively refining the process a model uses to form inferences by feeding it additional stored or real-time data. As a model takes in more data, the inferences should become more accurate: this is *learning*. Once inferences reach performance goals, the system can be put to practical use, inferring from new data. Notably, models are not necessarily fixed post-training; learning can continue after an AI model is put to practical use.

This section focuses on the dominant algorithmic technique for developing AI models—**machine learning**. Beyond machine learning, other methods are used to create AI models, including symbolic methods. Today, machine learning is the basis for most, if not all, modern systems. This technique is so dominant, in fact, that the term is largely synonymous with AI (box 6.1).

Symbolic methods are both an alternative and a complementary technique. Under symbolic AI, engineers try to build intelligence by treating knowledge as a collection of symbols—essentially core definitions, objects, labels, and representations that describe the world in human terms. "Cat," for instance, would be the symbol associated with the small, domesticated, feline commonly kept as a pet. Intelligence under this paradigm is a matter of constructing hierarchies and connections between these symbols and navigating those connections.[142] While less dominant, symbolic learning should not be ignored. Symbolic representations are essential to how humans understand the world, and these approaches continue to find use, often being paired with complementary machine learning systems.

Learning and Inference

The following are high-level illustrations of how machine learning and model inference work. In

BOX 6.1. Machine learning varieties

To create an AI system, engineers must select a machine learning algorithm. The algorithmic choice must be tailored to the task at hand. Although there is no one-size-fits-all strategy, most algorithms fall into one of the following categories:

1. **Supervised learning.** This approach follows a guess-and-check methodology. Data are fed into the model; the model forms a trial prediction (a guess) about those data, and, critically, that result is checked against engineer-provided labels, an answer key of sorts.[a] If the model's prediction differs from the correct label, the model then tweaks its processes to improve inference. Successive iterations thus improve performance over time. This method is useful for well-defined objectives and for situations requiring human terms and understanding. For example, supervised learning can teach algorithms to label images of fruit with their correct English name. It has also found use in content and model behavior regulation. GPT-4's engineers used supervised learning to fine-tune away certain undesirable behaviors and outputs.[b] Although useful for helping models understand data from a human perspective, this method's challenge is that models cannot learn what they are not trained to do. Their abilities are driven, restricted, and biased by the data chosen during the training process.
2. **Unsupervised learning.** Unsupervised learning algorithms are used when desired outcomes are unclear or broad. Unlike supervised learning, in which a system is trained to perform discrete and human-defined tasks, unsupervised learning models take in unlabeled data, sift through them, learn what hidden patterns and features

BOX 6.1. (*continued*)

they contain, and then cluster this information according to found categories and patterns.[c] This approach is useful in data analysis, where humans are prone to miss important data features and overlook unobvious correlations. Likewise, large language models such as GPT-3 often use unsupervised techniques to develop their broad set of skills and pattern recognition abilities.[d] Unsupervised learning benefits include looking at data through a detailed lens, doing so without many human biases and blind spots, and analyzing data with greater speed. Operating without a human-provided lens, however, can be a challenge. Although an unsupervised algorithm can categorize data and find patterns, it might not understand how to define its discoveries in human terms or match them to human objectives. Nonetheless, unsupervised techniques are behind many of the most cutting-edge systems in use today.

3. **Semi-supervised learning.** Semi-supervised learning is a hybrid of supervised and unsupervised learning that combines a portion of labeled data on top of a larger amount of unlabeled data.[e] This approach provides a light touch of supervision that can be helpful when some guidance is needed to direct the algorithm toward useful conclusions. It can be useful, for instance, when categorizing written text. The unsupervised half might first cluster the symbols based on their shapes. Then to label these groupings, the AI can learn their names using a human-provided answer key.[f] The result is an AI model that can recognize the alphabet.
4. **Reinforcement learning.** Reinforcement learning is driven by process rather than by data analysis. These algorithms use trial and error rather than **big data** to figure out the process behind a given task. To learn, an AI agent is placed in an environment and tasked with either maximizing some value or achieving some goal.[g] A driverless car might be tasked with minimizing travel distance between two points or maximizing fuel efficiency. The algorithm then learns through repetition and a reward signal. Through repeated trials, it tries a process and receives a reward signal if that process furthered its goal. It then adjusts its code accordingly to improve future trials.[h] This gamified approach is useful when a general goal is known, such as maximizing distance traveled, but the precise means of achieving that goal are unknown. **Reinforcement learning from human feedback**, for instance, is a prominent technique for aligning models with human preferences by testing them on users and rewarding them when users rate results positively. Here the goal of "usefulness for humans" is known, but how to get there and what that means is fuzzy—reinforcement learning's flexible style leans into that ambiguity. A major challenge with reinforcement learning is that sometimes AI can cheat by following strategies misaligned with human goals. For example, if the goal were to maximize fuel efficiency when navigating a group of naval vessels to a location, perhaps an AI might choose to destroy the slowest ships to increase total naval speed. Here the AI technically finds a more efficient process yet diverges from human intention.

In summary, supervised learning produces models that yield mappings between data, unsupervised learning produces models that yield classes and patterns in data, and reinforcement learning produces models that yield actions to take on the basis of data.[i]

NOTES

a. IBM, "What Is Supervised Learning?," accessed May 22, 2024, https://www.ibm.com/cloud/learn/supervised-learning.
b. OpenAI, "GPT-4 Technical Report," March 4, 2024, https://arxiv.org/pdf/2303.08774.
c. IBM, "What Is Unsupervised Learning?," accessed May 22, 2024, https://www.ibm.com/cloud/learn/unsupervised-learning.
d. Tom B. Brown et al., "Language Models Are Few-Shot Learners," Open AI, July 22, 2020, https://arxiv.org/pdf/2005.14165.
e. Jason Brownlee, "What Is Semi-Supervised Learning?," Machine Learning Mastery, April 9, 2021, https://machinelearningmastery.com/what-is-semi-supervised-learning/.
f. Ben Dickson, "What Is Semi-Supervised Machine Learning?," *TechTalks*, January 4, 2021, https://bdtechtalks.com/2021/01/04/semi-supervised-machine-learning.
g. Piyush Verma and Stelios Diamantidis, "What Is Reinforcement Learning?," Synopsys, updated April 27, 2021, https://www.synopsys.com/ai/what-is-reinforcement-learning.html.
h. M. Tim Jones, "Models for Machine Learning," IBM, December 5, 2017, https://developer.ibm.com/articles/cc-models-machine-learning/#reinforcement-learning.
i. Jones, "Models for Machine Learning."

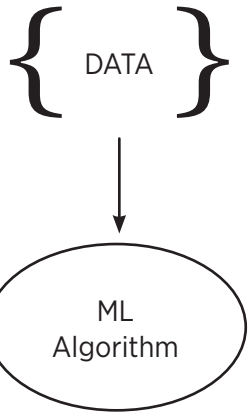
the Level 2 section, each of these is presented in a more detailed, yet still understandable, manner.

Learning

At a high level, how do AI systems learn? To illustrate this process, examine how a supervised learning algorithm builds its intelligence.

Fundamentally, this process starts with two elements (figure 6.1): data and the model one wants to train. To kick off the process, the as-yet unintelligent model will take in one piece of data from the dataset. Although it has not yet been refined in any way at this point, the model will then attempt an initial prediction based on that data. It does so to assess how well it performs so that improvements can be made.

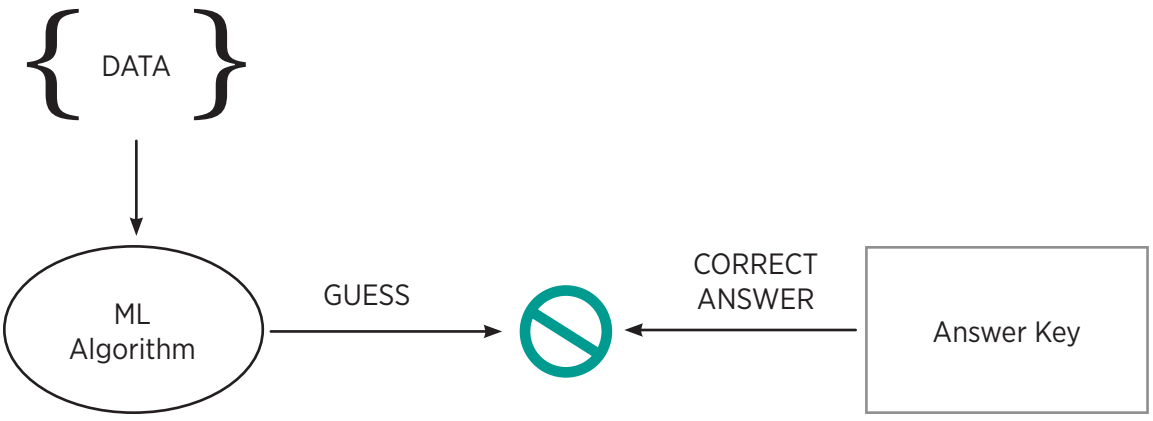
FIGURE 6.1. How artificial intelligence systems learn



Once this initial prediction is made, the model then needs a benchmark to score how well it performed. There are many types of benchmarks, but in the case of supervised learning, one can use an answer key of sorts (figure 6.2). Specifically, each data point will be given a human-provided label that represents the intended correct result. Suppose that the model is an image recognition system. If the training data included an image of an apple, it would be labeled with the correct term: "apple." If the model incorrectly produced the prediction "pear," the label would signal to the model that a mistake was made.

When the label and prediction differ, this incongruity signals to the model that it must change. Guided by a mathematical process, the model then gently tweaks certain internal settings and knobs called **parameters**, which are the values that shape its analytical processes. These tweaks ought to improve the model's

FIGURE 6.2. Benchmark model



predictive abilities for future trials. Note that although guided by mathematics, these tweaks do not guarantee improvement.

The algorithm repeats this process on the next piece of data. With each iteration, the model tweaks its parameters with the hope that collectively, these small changes allow it to converge on a state where it can consistently and accurately make high-quality predictions. Recall that proper training can require millions of data points and, by extension, countless rounds of training to converge on somewhat-reliable inferences.

Once the machine learning process is complete, the fully trained model can then be deployed and perform inference on real-world data that it has not seen before.

Inference

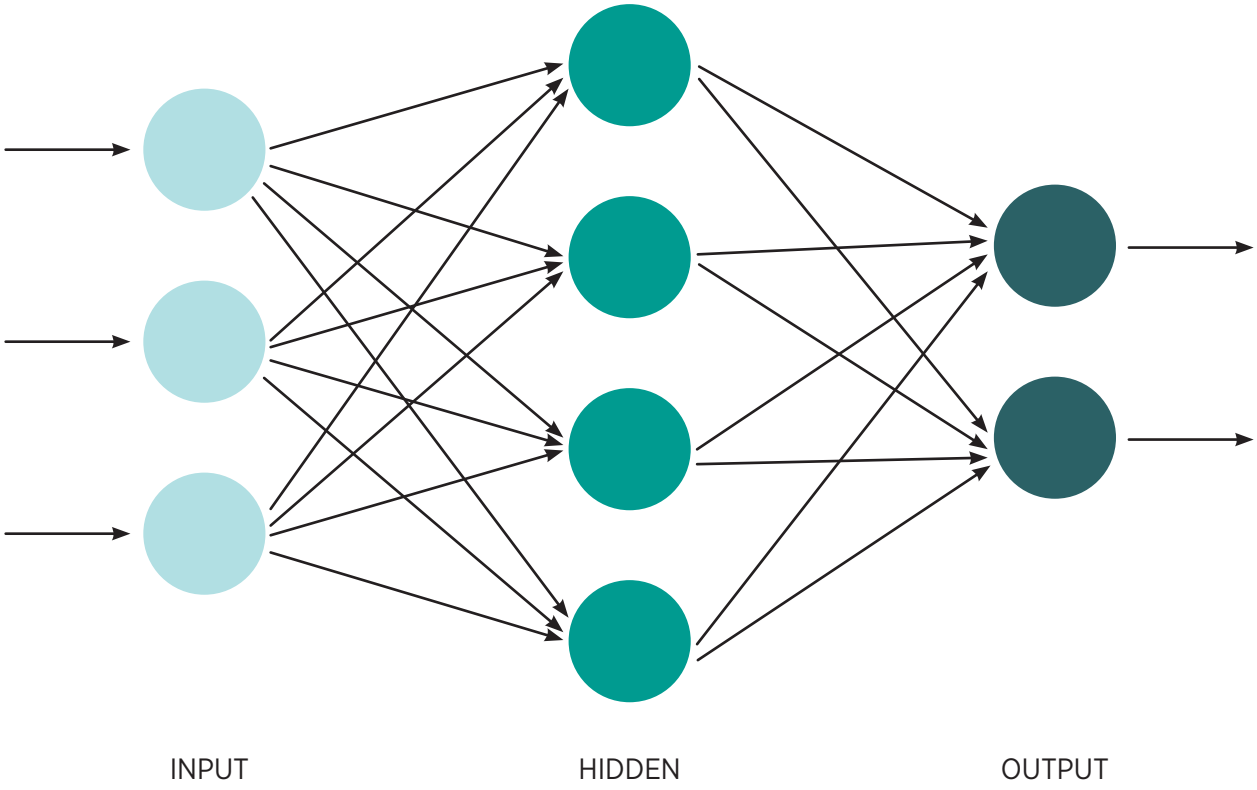
Once training is complete, how do these models perform inference on never-before-seen data? As is often the case, there are many tools that can be used. As an illustration, however, examine the most popular: the **artificial neural network** (ANN) (figure 6.3). This work uses neural networks to illustrate AI inference because such networks are behind most modern AI innovations, including driverless cars, image generation, AI-powered drug discovery, and large language models. Just as machine learning has become synonymous with AI, many observers often treat neural networks as synonymous with machine learning. Unlike the difference between machine learning and AI, however, other approaches are still widely used and very popular. Examples include regression models, which act to map the relationship between data variables; decision trees, which seek to establish branching patterns of logic that input data can follow to reach a conclusion; and clustering algorithms, which seek to sort data into clusters based on various metrics of data similarity.[143]

As the name implies, a neural network is an attempt to simulate the cognitive processes of the brain in digital form. These networks are composed of smaller units (the circles in figure 6.3) called **artificial neurons**. During the training process, each neuron will be tuned to find a unique and highly specific pattern in the input data that is highly correlative with accurate predictions. For instance, a neuron in a network designed to identify a face might be tuned to look for the visual patterns that represent a mouth, a pattern well correlated with faces. These patterns are the basis of the network's decisions.

To analyze a given piece of data, the network will first pass that data point into a set of neurons called the input layer. This is the far-left column in figure 6.3. Each neuron in this set will then examine the data for whichever patterns it has learned are significant. After this first round of analysis, these discovered patterns are then fired to downstream neurons.

When one neuron communicates with another, the information it sends is given a **weight**, which tells its neighboring neurons the importance of the pattern it has discovered for determining the final prediction of the network. Weighting certain patterns gives them an outsized influence on the final predictions. This approach is useful because it allows the network to prioritize what information is worth attention. If a network were trying to determine whether an image were a face, a freckle might receive a low weight because this feature is not highly indicative of a face; the freckle could be on an arm, a leg, or anywhere else. An eye, however, would receive an exceedingly high weight because this feature almost perfectly correlates with the prediction that an image is a face.[144] These weights are one

FIGURE 6.3. Artificial neural network



Note: Each dot represents an artificial neuron, and each arrow represents a connection between these neurons.

of the tunable **parameters** mentioned previously that are used to guide network analysis. Subsequent neurons take these weighted patterns and use them to find more complex patterns within patterns, developing an ever more nuanced picture of what the data represent. If two neurons have each identified an eye, these two features can be combined by a downstream neuron into the more complex and perhaps descriptive feature, "pair of eyes."

At the end of this process, all of the information will be passed to the output layer of neurons that is tasked with determining which prediction is best correlated with the total sum of discovered patterns. That prediction will be the final output that can be used for further decisions, actions, or analyses.

Before moving on, note the advantages of this structure. First, this format allows the system to divide and conquer. With hundreds, thousands, and sometimes millions of neurons deployed to look for specific, fine-grained patterns, networks can capture the deep nuance and complexity of real-world data. Dividing and conquering gives networks both flexibility and greater accuracy.

Second, the connections between neurons allow for discoveries to be shared and combined, deepening analysis. Individual patterns, on their own, are often not enough to properly predict

what data represent. By combining patterns through neuron-to-neuron communication, a neural network forms a more complete picture. To facilitate this, modern networks are often structured in **layers** of neurons, each of which takes in past patterns and recombines them in new and ever more complex ways. As a result, machine learning that uses neural networks is often referred to as **deep learning**,[145] a term that describes the multiple layers of neurons that data must pass through before a final prediction can be made.[146]

Generative AI

The recent explosion of AI interest largely centers on generative AI, systems trained to create high-quality text, media, or other data. Most generative AI systems today wield a shared **model architecture**, a deep learning design scheme that dictates how data interact with and flow through a model, called the **transformer**. What distinguishes the transformer is its ability to remember and connect disparate pieces of input data, rapidly process many data in parallel, and efficiently scale to learn and process a vast collection of data. The transformer has enabled the generative AI boom. This success, however, also rests on improvements in microchip processing power, specifically graphics processing units equipped to manage a transformer's often immense scale, and the **big data** needed to capture the diversity of knowledge that generative systems require. The transformer's flexibility has enabled immense breadth, finding application in image and video generators, voice cloning, music generation, machine translation, materials discovery, drug discovery, and other systems.

Large language models (LLMs)—generative models trained to understand, generate, and process human language—have received unique attention. LLMs include chatbots such as ChatGPT and Claude and machine translations systems. LLMs are also commonly integrated into other AI systems such as image and audio generation that require human language prompts. Likewise, language models are increasingly trending toward **multimodality**, or a model's ability to understand or produce multiple types of data, often including text, image, video, audio, and various computer file types.

A key LLM and generative AI concern is **hallucinations**, outputs that are incorrect, unrelated to the prompt, or inconsistent with reality. While work is under way to decrease hallucinations, a perfect solution that always guarantees correct results is unlikely.

Key Challenges

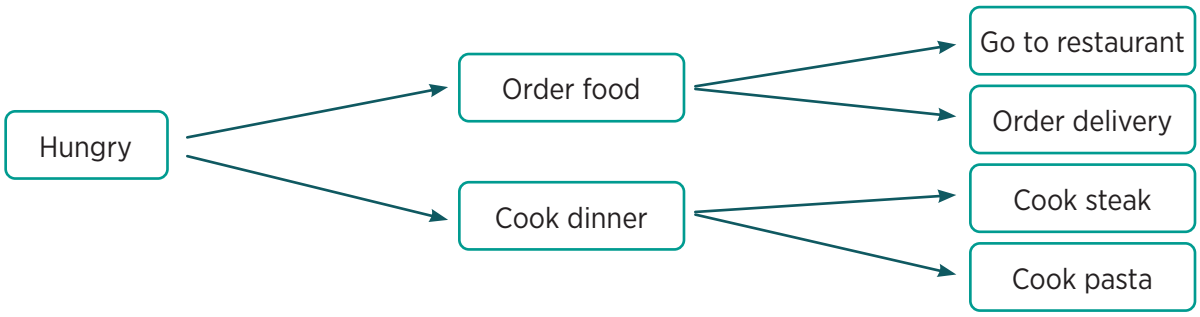
The key challenges of algorithms are model bias, explainability, and auditing of AI.

Model bias

As mentioned earlier, AI systems are not free from human biases. Although data are usually the root of many biased outcomes, model design is an often overlooked contributing factor. The frame of the problem that engineers are trying to solve with AI, for instance, naturally shapes how the model is coded.

For example, trying to design an AI system to predict creditworthiness naturally involves a decision on what creditworthiness means and what goal this decision will further.[147] The model's code will reflect this choice. If a firm simply wants to categorize data, perhaps a supervised learning algorithm can be used to categorize individuals. If the firm seeks to maximize profit,

FIGURE 6.4. Sample Decision Tree



Note: Figure 6.4 is a simplified sample output of how a decision-tree data algorithm might classify data by certain features it has learned during the training process.

perhaps a reinforcement learning algorithm could challenge the system to develop a process that maximizes returns. These differences in goals and model design decisions will naturally change outcomes and create qualitatively different AI systems. How a model is trained can also affect results. A model intended for multiple tasks has been found to show different outcomes when trained on each task separately rather than all at once.[148] Other such variations in the design process can be expected to yield varying results.

Mitigating this form of bias can be challenging and, like data bias, lacks a silver-bullet solution. Best practices are still developing, but suggestions tend to focus on process, emphasizing team diversity, stakeholder engagement, and interdisciplinary design teams.[149]

Explainability

Deep learning promotes large algorithms with opaque decision processes. Generally, as AI models balloon in size and complexity, explaining their decision-making processes grows difficult. Decisions that cannot be easily explained are called **black box** AI. Large neural networks, and their convoluted decision paths, tend to fall into this category. As a result, interest has grown in **explainable (or white box) AI**, a field that involves either designing inherently interpretable machine learning models whose decisions can be explained[150] or building tools that can explain AI systems.[151]

Some classes of **inherently interpretable** models exist today. For instance, decision trees—models that autonomously create "if–then" decision trees to categorize data—can be visually mapped for users (figure 6.4).[152]

Inherently interpretable models, however, are limited in accuracy and scale. Few modern neural networks are inherently interpretable, and model interpretation tools are an area of active research and development. Examples include tools that can determine what features in the input data were most significant in determining the model's conclusions.[153] The field is deeply underdeveloped, however, and cannot provide model-wide explanations, explain correlations between features, or produce necessarily understandable explanations.[154]

In many cases, applications of AI may require explainability. To abide by the law, an AI hiring system may need to prove that its decisions are not based on protected class characteristics. Explain-

ability can also help maximize policy effect. If decision makers know how a system produces decisions or results, that knowledge can enable targeted modifications of code to improve functionality or perhaps minimize unwanted biases.

AI auditing

Tangential to explainability is AI auditing. Given concerns over fairness, bias, correct design, and accuracy, there is significant interest in evaluating AI systems to ensure that they meet certain goals. Proposing AI audits, however, is easier said than done. Implementation naturally requires clarity of purpose. AI design challenges are rooted not only in technology but also in data, the application of technology, and social forces imprinted in these systems via biases. Choosing which problems to solve and what benchmarks to hit is an inherently messy task. As discussed earlier in this work, evaluation metrics and benchmarks are diverse and application specific.

At present, technical and ethical standards are fragmented, with little broad-based consensus. A 2021 Arizona State University study found an unwieldy 634 separate AI programs dedicated to developing soft law—that is, nongovernmental standards for AI development and governance.[155] This finding demonstrates that consensus has not been reached on the exact benchmarks and principles that might be used to audit AI.

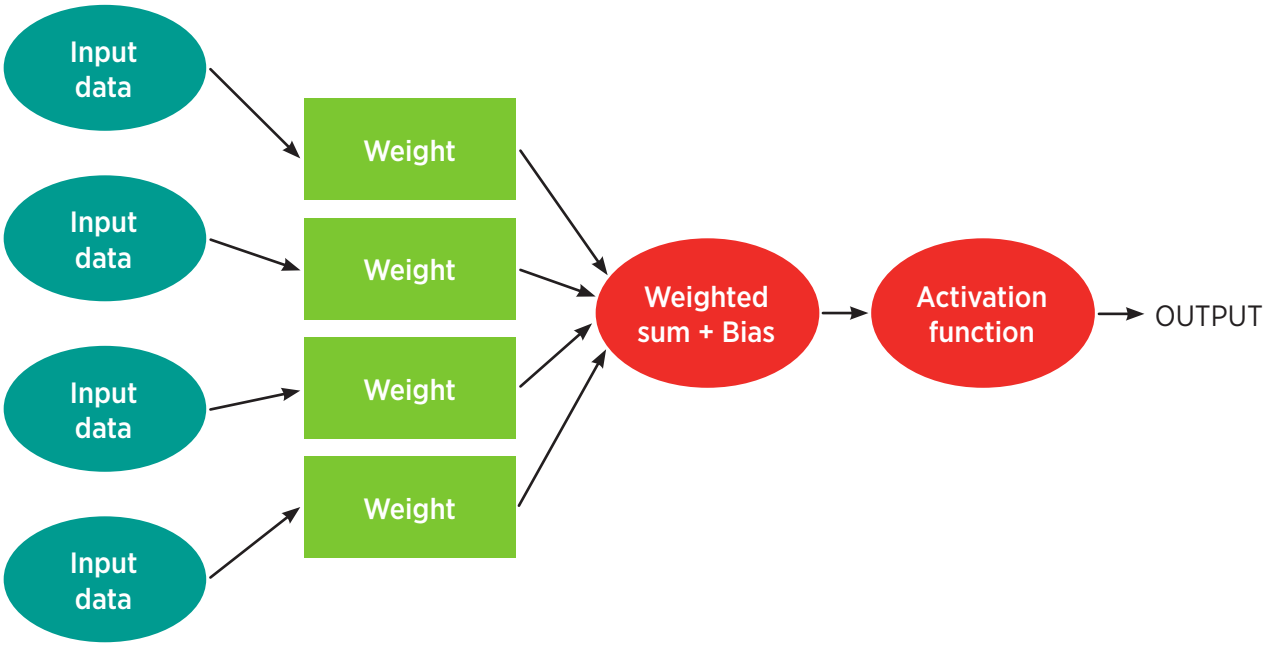
Process is another challenge. As a relatively new concept, AI audits lack frameworks and best practices, and commentators have noted that research on testing, evaluation, verification, and **validation** of AI algorithms has not kept pace with other subdomains of AI innovation.[156] Current processes and technologies offer no single audit technique that can test for the full range of possible errors.

Existing audits use a variety of methods. The data used to train algorithms can be audited to ensure that they are representative and avoid biases that might lead to disparate effect or to simply eliminate data extraneous to engineering goals. Black box testing, in which test data are fed into systems to analyze behavior, can help analyze general accuracy and stress test for certain undesirable biases. Model code can also be analyzed to better understand its process and its decision-making.[157] This method, however, is challenging because code is often complex and unwieldy, and the results of that code inherently depend on the inputs that are used.

As with all software, AI will be in a constant state of flux as updates are made and security patches released. Further, not all problems can be discovered through a single audit. Some challenges can be seen only once an AI is deployed in a complex human environment. To these ends, the National Institute of Standards and Technology (NIST) has proposed an iterative audit process that audits AI throughout its life cycle, during development and testing, and continuously after deployment.[158] Repeated scrutiny could help catch errors at each stage of the process and reinforce design principles to ensure that they are always top of mind. NIST's proposed process, however, is still in development. Best practices will require time and iteration before broad process agreement can be reached.

Each application naturally carries application-specific performance expectations. The issues faced by a medical AI system will differ from those of a music-generation AI.[159] Determining the questions that must be asked, the processes followed, and the issues to be tested will therefore require diverse thought and subject-matter expertise. As with the field, AI audits depend on application.

FIGURE 6.5. Model of a perceptron, a form of artificial neuron



DEEPER DIVE

Artificial Neurons

The previous section discussed machine learning and AI inference at a high level. This section discusses how an individual neuron might take in data and spot patterns within those data to produce good predictions. The general principles are illustrated by use of the common supervised learning process and the perceptron, a simple yet powerful artificial neuron model (figure 6.5).

Figure 6.5 is a diagram of an artificial neuron. On the left, the blue circles represent the input data for analysis. On the far right, the black arrow represents the final prediction that the model will output for the user. The core magic of this model, however, is the center. There one finds several elements that, while perhaps complex-looking at first, are relatively simple in operation.

An example follows.

Input

Start at the far left with the blue data inputs. For this example, suppose one operates a bank and is trying to train an algorithm to categorize loan applicants as either prime or subprime borrowers. Now suppose the applicants must submit four categories of data:

1. Whether they hold a savings account: represented by a 1 (yes), or a 0 (no)
2. Their number of dependents
3. Their number of monthly bank deposits
4. Their income bracket (represented by 1–7, with 7 being the highest)

For this illustration, suppose that the loan application for the neuron to analyze is as follows:

1. Savings account: 1
2. Number of dependents: 0

3. Number of monthly deposits: 2
4. Income bracket: 7

Data adjustment and activation

Detecting patterns in data is actually a process of transforming input data into an output that represents a meaningful pattern. This is done in two steps. First, the neuron manipulates the input data to amplify the most important information and sums the data. Next, it passes this sum to an **activation function**. In a realistic sense, the activation function represents the rules that transform the input data into the output decision. In many cases, however, it can more or less be thought of as an algorithmic trigger that needs to be tripped for the neuron to activate.[160] The activation function compares the manipulated data to certain criteria, which dictate the final output that the neuron will produce. In the simple prime or subprime case, this criterion is a threshold number: if the sum is higher than this threshold, the neuron sends a result indicating that this is a prime borrower. If not, it indicates subprime. Although in this case, the result is the neuron's final decision, note that in complex neural networks, the result might just be one of many patterns identified in service of the final decision.

Elements of an artificial neuron

Next, examine the tools that this neuron uses to adjust the data and calculate the final result. Surprisingly, this can be quite simple. In many cases, the math involved uses only simple arithmetic.

Once the data enter the neuron, they encounter the green squares in figure 6.5; each represents a **weight**. Using weights, the neuron can amplify a certain element of the input data through multiplication. For instance, it is likely that the income bracket data in this example are strongly correlated with prime borrowers; therefore, this feature of the data should be amplified in the final decision. To do so, one multiplies that value by a weight to make it bigger, giving it more significance.

Weights are a useful tool because they allow the truly important elements of the data to have an outsized effect on the result. Crucially, weights are a **parameter** that can also be tuned. The more important the value, the bigger a weight multiplier it will receive. Conversely, unimportant data can be eliminated by multiplying them by 0. Finding the correct weightings of data values can be seen as one of the core elements of a neuron's intelligence.

After the data have been weighted, they are added to a **bias value**. The bias acts as the threshold, mentioned previously, that the weighted data must surpass for the neuron to activate. Put another way, the bias puts a thumb on the scale of the result by adjusting what causes the neuron to trigger.[161] For instance, if prime borrowers should be rare, one might subtract a bias value, making it harder for the summed weighted data to trip the activation function.

After the data have been adjusted, they are then fed to the activation function. In the example neuron's case, if the final value adds up to 1 or greater, the neuron communicates a prime result; if not, it indicates subprime.

Calculation of the result

Let's assemble each element to see how it affects the data. As mentioned earlier, to produce a result, the neuron will simply take the input data—the loan application—multiply each category by its weight, and add these results with the bias value.

In this case, start by weighting the data. The data values are in blue, their weights in green, the bias in purple, and their sum in red:

> Result = 2 * (savings account) + 10 * (number of dependents) + 3 * (number of monthly deposits) + 1 * (income bracket) – 15.

Each data category is multiplied by a weight consistent with the importance of that data element in making final predictions. Run the data through this equation:

> 2 * (1) + 10 * (0) + 3 * (2) + 1 * (7)

The weighted data sum is 15.

Next, add the bias. Remember that the bias is essentially the threshold that the data need to surpass for the neuron to activate. According to the rules prescribed by the activation function, these values must be greater than or equal to 1 for the neuron to indicate a prime value. The result is in red, the weighted sum from the previous step is in black, and the bias is in purple:

> **Result** = 15 – 15

The result is 0. Therefore, the neuron chooses to categorize the data as subprime.

The learning process

For the sake of illustration, suppose that the model is currently in training and this result is not correct. The original data show that the individual is in the highest tax bracket and likely a prime borrower, yet the model in its current form classified the person as subprime. Thankfully, machine learning algorithms can learn from their mistakes and revise their weights and biases to produce better predictive outcomes.

How might this work? First, the algorithm must realize there was a mistake. In supervised learning, to train a model, engineers will use a dedicated set of **training data**[162] paired with labels that act as an answer key. In this case, the model will compare its result to the key and find that it made a mistake. This result will prompt the algorithm to adjust its parameters.

These changes are often made using educated guesses, guided by mathematics. There are a variety of methods, but usually the algorithm will base its actions on how much its prediction diverged from the correct answer. This is called the **loss**. That value is then used to adjust each of the weights up or down depending on whether they are causing the neuron to undershoot or overshoot the correct result. The goal is to minimize this loss value in future iterations.[163]

For the sake of simplicity and sanity, the somewhat complicated linear algebra involved here is not discussed. The key takeaway is that to improve, the algorithm adjusts its weights based on how much it erred, nudging the model in the direction of the correct answer. Each adjustment is not perfect but a mere educated guess. After enough trials, however, the process helps minimize loss and optimizes the algorithm.

Back to the example, suppose the model has subsequently altered its weights to make better predictions:

> Result = 2 * (savings account) + 1 * (number of dependents) + 3 * (number of monthly deposits) + 10 * (income bracket) – 15

Using *this* equation, the data would produce a result of 63. This is obviously greater than 15,

the threshold that the results must surpass for the activation function to signal a prime result. The model has now learned when to classify this individual as a prime borrower.

Training considerations

Once a network is properly trained, its results are tested using a dedicated set of **test data**. This test set includes unused data to assess accuracy and flexibility. Test data help avoid the problem of **overfitting**, a situation in which a model is tuned so precisely to the training data that it cannot adequately account for unexpected variations in new data. The opposite problem is the challenge of **underfitting**, a situation in which the model has not been properly tuned to the problem because of poor data or design, and accuracy suffers. Both can be detected using test data. When designing models, engineers must strike a balance between overfitting and underfitting.

Model tuning

Recent research suggests that adding greater depth and more neurons does not diminish returns on predictive performance.[164] That said, simply building increasingly massive models is not always feasible, given limitations in computing power. Model designers must therefore size their models to fit the data and computational power at their disposal. For instance, a programmer with just a simple laptop CPU wouldn't be able to design a model with hundreds of thousands of neurons. Insufficient data also constrain model size. The bigger the model, the more data it will need to be well tuned. If engineers do not have enough data, they would choose model alternatives that are smaller and differently resourced.

Beyond the size and scope of models, engineers also work to tune a model's **hyperparameters**, the settings that control the model's function.[165] An example of a hyperparameter is the learning rate. This rate dictates how large the tweaks to the model's weights will be each time it makes an adjustment. A higher learning rate increases training speed, at the cost of accuracy, and a lower training rate decreases training speed, with accuracy gains. The chosen settings, as with model size, depend on the engineer's specific resources and goals.

Finally, the engineer must also choose the correct model. Not all models are equal, and each comes with different strengths. Engineers must choose the best model for their goals. If a model for a given task does not exist, engineers can of course develop their own. That said, the majority of machine learning engineering relies on prefab models found in numerous **libraries**, many of which are free and open source. For example, the scikit-learn library includes a multitude of models that can be freely used and implemented using the Python programming language.[166]

Note that most AI engineering is unscientific. Rules of thumb have come to dominate AI. There are no set rules that govern the specific number of neurons required, for instance. This adds further bias to AI. These algorithms, much like data, are reflections of the skill and goals of engineers. The systems are not perfect, nor are they scientific. They can, however, still produce highly accurate results.

Model variety

The neuron illustration specifically presented a **feed-forward neural network**, a classic form that takes in data and directly maps them to a specific output.[167] For the prime/subprime categori-

zation task, this process worked perfectly. However, not all tasks are quite so straightforward. Some data, such as text, depend on complex relationships. The placement of a given word in a sentence depends not only on the words before it but also on those that follow. Analyzing a sentence requires a network that can analyze each word sequentially *and* keep track of how each word fits into the context of the sentence. Even more complexity enters the picture when neural networks are applied to generative tasks—that is, when they are asked to produce text, paint pictures, write songs, and so forth. These complex tasks are not simple categorization exercises. As such, numerous tools and models have been developed to augment the basic neural network structure and account for the unique complexities that come with each type of task.

The following is a short list of some of the dominant forms of neural networks and the tools used by these networks to produce high-quality results. Given the dynamism of the field, this list cannot detail all types and combinations of neural networks, nor can it predict which may fall out of favor.

Diffusion models. These models learn to generate new data, such as images, by gradually removing noise from a random input. Imagine starting with a blank canvas and randomly splattering paint all over it. The diffusion model learns to slowly and carefully remove the random splatters, step by step, until a clear, recognizable image emerges. It does this by training on a large dataset and learning the patterns and structures present in the data. The model can then apply this learned knowledge to generate new, previously unseen data similar to the training data. As of 2024, diffusion models have shown state-of-the-art results in generating highly realistic images, video, and audio.

Generative adversarial networks (GANs). A GAN is a training model that uses two separate neural networks that compete against each other to learn and improve. One produces fake data trying to trick the other into misclassifying them as real, while the other is competing to improve its abilities at distinguishing the fake data from the real data. This process creates an arms race of sorts, with both models adjusting themselves to improve their ability to produce fake data that look real and their ability to distinguish real from fake, respectively.[168] Theoretically, both models improve, and this refinement results in the ability to produce high-quality artificial data. This method is widely useful in applications in which unique data must be generated, including AI-created art, images, video, and deep fakes.

Convolutional neural networks (CNNs). CNNs are neural networks used in image and video analysis. These models uniquely use convolutional layers, which act as data filters trained to spot and separate patterns that are highly correlated with a specific result. The result from these layers simplifies data and accentuates the most important features.[169] For example, if an algorithm is trained to recognize dogs in images, a convolutional layer may be trained to specifically find the pixel data patterns that form floppy ears. If this layer spots this pattern, there is a high likelihood that the image is indeed a dog. Overall, these layers act to break down images into their component patterns and unlock greater predictive powers for neural nets.

Recurrent neural networks (RNNs). These networks are defined by their ability to "remember."[170]

As data flow through an RNN, they are analyzed on their own merits, and their qualities are knit together and compared to the data that came before, allowing the network to see patterns over time. This temporal analysis quality has applications in time-dependent data such as video or writing.

Transformers. An architecture that arrived in 2017, it has since been widely applied across many complex tasks such as natural language processing. Transformers' key selling point is their **attention mechanism**, which allows the model to "pay attention" to key features and remember how those features in the data relate to others.[171] This quality allows these models to treat data as a complex whole, a characteristic that is essential for any task that requires understanding over time, such as reading text. Because the transformer is an architecture, it is often used in concert with other models. OpenAI's Sora video generator, for instance, is a diffusion transformer using both diffusion models and the transformer architecture. The basis for many **foundation models** today is the transformer.

7. Conclusion: The Policymaker's Challenge

While the goal of this introduction to artificial intelligence (AI) is simplicity, some may find the staggering breadth of AI unwieldy. AI's wide scope is a natural consequence of its general and often ill-defined nature. Recall that, fundamentally, AI is a normative goal. As with any goal, it can be defined in a variety of ways depending on the user and the context. One goal might be to wield and design AI systems to maximize safety, another might involve a minimizing bias, and third perhaps would prioritize national security. Such general goals grow more specific and varied as systems are designed and applied in application-specific contexts.

The fundamental challenge for policymakers will be recognizing this diversity and understanding that not all AI goals will coexist peacefully, nor will they necessarily match the goals of policymakers. Any regulation or AI-related policy will naturally involve a normative choice. What *should* AI look like, what *should* it do, and how *should* it be used—that is, what goal or set of goals are encouraged or allowed?

Diversity is perhaps the best first step toward meeting this difficult challenge. Only through application- and sector-specific knowledge can the full range of potential AI goals, applications, and issues be understood. Meeting the challenge of AI's increasing breadth requires a representative breadth of policymakers to understand AI. This general-purpose technology is also a general-purpose policy issue.

Having peeked under the AI hood, readers should have a technical starting point that can be customized and applied to each given sector and field. Today, AI systems are changing—and perhaps even transforming—many fields. With such potential, it is incumbent on all policymakers to dig in, understand these concepts, and grapple with the diversity of these impactful systems.

Glossary

Accuracy: An evaluation metric that measures the reliability of a system's inferences.

Activation function: The mathematical function that transforms data inputs into outputs. This shapes the final predictions that are made and serves as the algorithmic trigger that needs to be tripped by input data for a given prediction to be made.

Adversarial examples: Data inputs maliciously designed to trick AI systems during inference.

Adversarial machine learning: Refers generally to the study and design of machine learning cyberattacks and defenses.

AI alignment: In the context of artificial general intelligence, alignment of AI systems refers to their correspondence with generally accepted human values (do not harm, do not kill, protect the vulnerable, allocate human rights equitably, and so on).

AI chips or **AI accelerators:** A range of chips designed specifically for the unique processing needs of AI.

AI triad: The three primary input technologies that yield artificial intelligence: microchips, data, and algorithms.

Algorithm: A logical sequence of steps to accomplish a task, such as solving a problem.

Alignment imbalance: A state in which AI is generally misaligned with human values. This imbalance supposes that AI systems can possibly be balanced with human values. However, imbalance may be inherent to all AI systems and baked into their design.

Application-specific integrated circuits (ASICs): The fastest and least flexible form of AI chip. ASICS are single-purpose chips and cannot be rewritten; the algorithms they use are hard wired into their silicon.

Artificial general intelligence: A general-purpose AI system that can adapt and learn any task. It is not designed for a specific narrow purpose or set of purposes.

Artificial intelligence (AI): The goal of automating tasks normally performed by humans. To reach this goal, one uses "machine-based system[s] that can, for a given set of human-defined objectives, make predictions, recommendations or decisions influencing real or virtual environments."[172]

Artificial narrow intelligence: AI built for a narrow purpose, such as a specific application. This AI can do one or a few tasks and do so with high accuracy, but it cannot transfer to other applications outside of its design mandate.

Artificial neural network (ANN): A type of model formed from networks of interconnected artificial neurons. Neurons take in data, divide that data, and parse these divisions to discover patterns. Patterns are then assembled to form increasingly advanced patterns and ultimately inform the network's final predictions.

Artificial neurons: Individual components of ANNs that take in data and look for specific patterns in such data that they have learned are significant during the training process.

Attention mechanism: A component of certain neural networks which allows the model to "pay attention" to key features in data and remember how those features in the data relate to others.

Bayesian methods: Models that are coded with previous information that provides context and shrinks the overall learning task and, by extension, the required training data.

Benchmarks: Common datasets paired with evaluation metrics that can allow researchers to compare the quality of models.

Bias: Defined generally as the difference between desired outcomes and measured outcomes. Often it refers to human biases inherited in AI systems through model or data design choices.

Bias value: The threshold that the weighted data must surpass for a neuron to activate. Mathematically, this serves as the intercept that orients the activation function toward the "shape" of reality.

Big data: Big data AI systems are trained on large, representative, and diverse datasets that are expected to capture all the corner cases and details of a given problem. The theory is that by training an AI system on such a dataset, the system should, it is hoped, capture and learn all the required details of a given problem.

Binary: A numerical system that represents values in series of just 1s and 0s. Most data in computer science and AI are represented in this form.

Bit: The smallest unit of data that represents a binary choice between a 1 and a 0.

Black box: The often-opaque decision-making processes behind deep neural networks.

Byte: A data unit the size of 8 bits.

Central processing unit (CPU): A type of general-purpose chip designed to handle all standard computation.

Chatbot: A class of AI technology that takes language-based prompts and responds in a language or text-driven, often conversational manner. Examples include ChatGPT and even AI

assistants such as Amazon's Alexa. Chatbots have been around for decades; however, today's chatbots often wield large language models (LLMs), such as GPT-4, to ensure advanced flexibility, fluidity, and generalization. For sensitive or regulated applications, such as finance, more stringent technologies such as inflexible decision trees remain common.

Circuits: Electronic components linked together to enable certain computational functions such as addition, subtraction, or memory storage.

Cloud computing: A general computing concept in which computing resources (both memory and processors) are stored remotely.

Code: The set of instructions given to a computer system.

Computer program or **software:** Code for the operation of a computer application.

Convolutional neural network (CNN): A form of neural network that uses convolutional layers, which act as data filters trained to spot and separate patterns that are highly correlated with a specific result. These layers simplify data and accentuate the most important features. CNNs can be useful in many applications, such as image analysis, financial time series analysis, and natural language processing.

Data: In the context of computer science, data are pieces of discrete information that can be encoded, stored, and computed.

Data cleaning: The process by which data are prepared for use by an AI algorithm.

Data poisoning attacks: Attacks on AI systems caused by the malicious manipulation of data.

Data standards: Industry and application-specific standards that dictate in certain circumstances what data must be recorded and how that data must be recorded.

Data warehouses: Large, centralized warehouses holding hundreds of servers on which vast lakes of data are stored and large-scale computations are run.

Deep learning: A type of machine learning that specifically uses deep, multilayered neural networks.

Deposition: A process used in chip fabrication that blankets chips with materials to add components.

Dopants: Intentional impurities that lace the silicon in transistors, changing when and how transistors switch between conducting or insulating electric current.

Electronic design automation (EDA): The software used by hardware engineers to design computer systems and chips.

Etching: A process used in chip fabrication that uses chemicals to remove unwanted material and shape the design of the chip.

Evaluation metrics: Metrics that can be used to assess AI system quality. These are diverse, and the metrics selected should match application needs and engineering goals.

Execution units: Microprocessor subsystems that package related circuits together with memory and other tools to enable basic functions.

Explainable or **white box AI:** An emerging class of AI that seeks to provide explanations of how the system's decisions and predictions are made.

F1 score: An evaluation metric that assesses how well a model minimizes both false negatives and false positives.

Feed-forward neural network: A type of machine learning in which data flow in one direction through the network's layers.

Federated learning: A training technique that trains AI models on a web of disconnected servers or processors, rather than a centralized server, often to eliminate data aggregation and preserve privacy.

Few-shot learning: The ability of a model to form accurate inferences trained on only a few explicit examples of the problem at hand.

Field-programmable gate arrays (FPGAs): Task-specific chips that can be written and rewritten for a single-purpose algorithm. Given their task specificity, FPGAs are faster than GPUs. They are still slower than ASICs, because their ability to be rewritten comes with certain speed costs.

File format: A type of data standard that defines how data are digitally represented.

Fine-tuning: The process of refining a general-purpose foundation model toward specific goals or tasks. Fine-turning often involves additional training on task- or domain-specific data, training in controls to limit certain undesirable behaviors, or further training to align models toward certain desired behaviors.

Floating point operations per second (FLOPS): A measure of computational speed and performance that clocks the floating point operations, the number of mathematical operations a processor can complete in a second. Confusingly, floating point operations (FLOPs with a lowercase "s") are also used to measure model size based on how many operations that model requires.

Foundation models: Large-scale machine learning models trained on broad sets of data that can be easily adapted to a wide range of downstream tasks.

Generalization: A system's ability to "adapt properly to new, previously unseen data."[173] Generalization is highly desirable and a marker of AI quality.

General-purpose technology: Innovations that "[have] the potential to affect the entire economic system."[174]

Generative adversarial networks (GANs): A form of neural network in which competing agents seek to outcompete each other. Through competition, each party improves, ultimately improving its overall predictive qualities. GANs are noted for their generative modeling, or creative, abilities. This specifically means that they use pattern recognition to predict how to best generate novel output content, such as images.

Generative AI: AI systems trained to create high-quality text, media, or other data. Generative AI is not limited to media. Protein folding systems, materials discovery systems, code generation, and other science, technology, engineering, and mathematics (STEM) applications can be considered generative AI.

Graphics processing units (GPUs): Limited-purpose processors that were originally designed for graphics processing but that have been reappropriated for AI. GPUs excel at matrix multiplication, a function central to AI, giving them speed advantage over traditional CPUs.

Hallucinations: Generative AI outputs that are incorrect, unrelated to the prompt, or inconsistent with reality.

Hyperparameters: High-level settings that can be adjusted by engineers to control the model's functions.

Inference: A probabilistic guess made by an AI system on the basis of patterns or trends observed in data.

Inherently interpretable: Models that by design are simple to interpret or understand.

Integrated circuits (ICs) or **microprocessors:** Devices that can perform basic operations of software commands.

Internet of things (IoT): Networks of diverse internet-connected devices. IoT devices often act as key data inputs to AI systems.

Large language models (LLMs): Generative models trained to understand, generate, and process human language. Machine translation and chatbots are common LLMs. To enable prompting, LLMs can also be integrated into systems such as image generators.

Layers: Collections of neurons that data must pass through simultaneously in a network.

Libraries: Databases of functions that can be plugged into computer programs. There are many free-to-use libraries of machine learning models that are commonly appropriated for AI.

Loss: In machine learning, this is the mathematical difference between the correct outcome and the desired outcome.

Machine learning: A method for iteratively refining the process a model uses to form inferences through feeding it stored or real-time data.

Memory units: Devices that use transistors and other components to store information. Memory units can be subcomponents of a chip or standalone chips depending on their size and function.

Microchip architecture: The "blueprint" configuration of chip components, including circuits, execution units, and input/output devices. AI chips depend on architectural changes for performance gains.

Model: The software configuration that results from machine learning. Once fed new data, the model can produce inferences in the form of predictions, decisions, and other outputs.[175]

Model architecture: An AI model design scheme that dictates how data interact with and flow through a model.

Moore's law: An observation stating that the number of transistors per chip doubles roughly every two years. More than an empirical observation, it was an expectation that came to organize the efforts of the microchip industry and was a self-fulfilling prophecy for a long time.

Multimodality: The ability of a model to understand multiple types of data, often including text, image, audio, and various computer file types.

Overfitting: A situation in which a model is tuned so precisely to the training data that it cannot adequately account for new data.

Parallelism: The ability of a chip to perform certain functions in parallel rather than sequentially, allowing faster processing.

Parameters: The values that shape a model's analytical processes.

Photolithography: A process used in chip fabrication by which light is shined through a "circuit stencil" known as a photomask, printing the design onto the chip's wafer.

Precision: An evaluation metric that evaluates how many positive results are true positives.

Recall: An evaluation metric that states the percentage of a model's negative results that are true negatives.

Recurrent neural networks (RNNs): Neural networks defined by their ability to remember past information and connect that information to future data. This "memory" is necessary in complex, time-dependent data such as video analysis, natural language processing, and other applications.

Reinforcement learning: A type of machine learning that uses trial and error to learn the best process to achieve a given goal. To learn, an AI model is given a scenario and tasked with maximizing a reward or achieving a goal. When its process improves, it receives a rewards signal that instructs it to reinforce the processes that led to that improvement.

Reinforcement learning from human feedback (RLHF): A prominent fine-tuning technique geared at aligning models with human preferences. During RLHF, systems are tested on or produce outputs for human users; when those users react positively, a reward signal is sent to the system, thereby helping it improve its outputs.

Representation: The concept of translating observable objects (images, words, sounds) into digital code.

Semiconductor devices: A class of devices that uses the unique switching properties of semiconductor materials to alert the flow of electricity. Example devices include LEDs and transistors. Microchips, ICs, and microprocessors are all made of semiconductor materials.

Semiconductor materials: Materials such as silicon that can act as either insulators or conductors of electricity.

Semi-supervised learning: A hybrid of unsupervised and supervised learning in which a portion of labeled data are provided to the model on top of a larger amount of unlabeled data. This approach can provide a light touch of supervision.

Small data: An alternative strategy to big data approaches that uses a variety of techniques to train AI algorithms on smaller datasets when information is poor, lacking, or unavailable.

Stale data: Outdated data that are no longer representative of a given problem.

Stochastic parrots: A term that describes AI systems that randomly rearrange and regurgitate learned data rather than provide true insight or understanding.

Superintelligence: An AI system that is smarter than humans in almost every domain.

Supervised learning: A type of machine learning that uses a guess-and-check methodology by which the model takes in data, makes a prediction about those data, and compares that prediction to a labeled answer key. If the inference is incorrect, the algorithm adjusts itself to improve performance.

Symbolic methods: An alternative and complementary technique to machine learning. Under symbolic AI, engineers try to build intel-

ligence by treating knowledge as a collection of symbols—essentially core definitions, objects, labels, and representations that describe the world in human terms.

Synthetic data: Data that are artificially created either by human or machine generation but still thought to be generally representative of a problem. Training AI on artificial data can supplement real-world data when quality data resources are limited.

Test data: The unique set of data reserved for testing the model for final accuracy and effectiveness used in machine learning. Test data must be separate from the training data.

Three V's: Key characteristics that define the quality of a dataset. *Variety* refers to the diversity of the data. *Volume* refers to the size of the dataset. *Velocity* refers to the usability and speed by which the data can be applied. Other publications may list four, five, or even six Vs. The term tends to vary depending on context and purpose.

Training: The process by which models take in stored or real-time data to refine their processes and improve their inferences.

Training data: The unique set of data reserved for the model training process in machine learning.

Transfer learning: One small-data approach that allows models to inherit learning from previously trained big data models.

Transformers: An emerging class of neural networks that uses a so-called attention mechanism that allows the model to pay attention to key features and remember how those features in the data relate to others.

Transistor: A device built from a combination of silicon and dopants, impurities that alter the properties of conductivity to provide discrete control by engineers over electric currents.

Underfitting: A situation in which a model has not been properly tuned to the problem because of poor design or data quality.

Unsupervised learning: A type of machine learning that focuses on sorting unlabeled, unsorted data and discovering patterns in those data. This method does not focus on specific outcomes but rather on discovering the meaning and patterns in data.

Validation: The process by which the engineer uses a dedicated validation dataset to tune the hyperparameters of the model. Generally, this is done after training but before testing.

Validation data: The unique set of data used during machine learning validation. These data are used specifically to tune the model's hyperparameters.

Wafer: The thin disk of semiconductor materials that acts as the base of a computer chip.

Weight: A numerical value that amplifies or suppresses the importance of a pattern found in data.

Zero-shot learning: The ability of a model to form accurate inferences without having been trained on explicit examples of the problem at hand.

Notes

1. Rishi Bommasani et al., "On the Opportunities and Risks of Foundation Models," Center for Research on Foundation Models at the Stanford Institute for Human-Centered Artificial Intelligence, revised July 12, 2022, https://doi.org/10.48550/arXiv.2108.07258.
2. Bommasani et al., "On the Opportunities."
3. Hayden Field, "The First Fully A.I.-Generated Drug Enters Clinical Trials in Human Patients," *CNBC*, June 29, 2023.
4. Remi Lam, "GraphCast: AI Model for Faster and More Accurate Global Weather Forecasting," *Google DeepMind Blog*, November 14, 2023, https://deepmind.google/discover/blog/graphcast-ai-model-for-faster-and-more-accurate-global-weather-forecasting/.
5. See the ChatGPT website at https://openai.com/index/chatgpt/.
6. James Pethokoukis, "How AI Is Like That Other General Purpose Technology, Electricity," *AEIdeas*, November 25, 2019, https://www.aei.org/economics/how-ai-is-like-that-other-general-purpose-technology-electricity; Elhanan Helpman, ed., *General Purpose Technologies and Economic Growth* (Cambridge, MA: MIT Press, 2003), https://mitpress.mit.edu/9780262514682/general-purpose-technologies-and-economic-growth.
7. Emily M. Bender et al., "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?," *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (New York: Association for Computing Machinery, 2021), 610–23, https://doi.org/10.1145/3442188.3445922.
8. Jacob Helberg, *The War of Wires: Technology and the Global Struggle for Power* (New York: Avid Reader Press/Simon & Schuster, 2022), 60–61.
9. "Huge 'Foundation Models' Are Turbo-Charging AI Progress," *The Economist*, June 11, 2022, https://www.economist.com/interactive/briefing/2022/06/11/huge-foundation-models-are-turbo-charging-ai-progress.
10. Melissa Heikkilä, "This Artist Is Dominating AI-Generated Art. And He's Not Happy about It," *MIT Technology Review*, September 16, 2022, https://www.technologyreview.com/2022/09/16/1059598/this-artist-is-dominating-ai-generated-art-and-hes-not-happy-about-it.
11. Natasha Lomas, "Glaze Protects Art from Prying AIs," *TechCrunch*, March 17, 2023, https://techcrunch.com/2023/03/17/glaze-generative-ai-art-style-mimicry-protection.
12. Rob Salkowitz, "AI Is Coming for Commercial Art Jobs. Can It Be Stopped?" *Forbes*, September 16, 2022, https://www.forbes.com/sites/robsalkowitz/2022/09/16/ai-is-coming-for-commercial-art-jobs-can-it-be-stopped.
13. National Security Commission on Artificial Intelligence, "2021 Final Report," March 2021, https://www.nscai.gov/2021-final-report.
14. James E. Baker, *The Centaur's Dilemma: National Security Law for the Coming AI Revolution* (Washington, DC: Brookings Institution Press, 2020).

15. Shane Legg and Marcus Hutter, "Universal Intelligence: A Definition of Machine Intelligence," December 20, 2007, https://doi.org/10.48550/arXiv.0712.3329.
16. National Artificial Intelligence Initiative Act of 2020, Pub. L. No. H.R. 6216 (2020).
17. François Chollet, *Deep Learning with Python*, 2nd ed. (Shelter Island, NY: Manning Publications, 2021).
18. Alexandre Gonfalonieri, "How Amazon Alexa Works? Your Guide to Natural Language Processing (AI)," *Medium*, November 21, 2018, https://towardsdatascience.com/how-amazon-alexa-works-your-guide-to-natural-language-processing-ai-7506004709d3.
19. Ben Buchanan, "The AI Triad and What It Means for National Security Strategy," *Center for Security and Emerging Technology* (blog), August 2020.
20. Buchanan, "The AI Triad."
21. This new class of automation has raised many questions of ethics, safety, and the role of government, all of which have limited autonomous system deployment. Driverless car engineers often puzzle over how driverless vehicles should confront classic trolley-problem scenarios. AI safety experts often worry about determining acceptable levels of failure before deploying these systems. Meanwhile, new forms of automation can be blocked by historical laws and regulations written under the assumption of human, not machine, control.
22. Narrow AI is often alternatively referred to as "brittle."
23. Ariel Conn, "Benefits and Risks of Artificial Intelligence," Future of Life Institute, November 14, 2015.
24. Google Developer Program, "Machine Learning: Foundational Courses, Generalization," accessed May 20, 2024, https://developers.google.com/machine-learning/crash-course/generalization/video-lecture#:~:text=Generalization%20refers%20to%20your%20model's,Time%3A%205%20minutes%20Learning%20Objectives.
25. McKenna Fitzgerald et al.,"2020 Survey of Artificial General Intelligence Projects for Ethics, Risk, and Policy," Global Catastrophic Risk Institute, December 31, 2020, https://gcrinstitute.org/2020-survey-of-artificial-general-intelligence-projects-for-ethics-risk-and-policy.
26. Fitzgerald et al., "2020 Survey of Artificial General Intelligence Projects."
27. Fitzgerald et al., "2020 Survey of Artificial General Intelligence Projects."
28. Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies*, 1st ed. (Oxford, UK: Oxford University Press, 2014).
29. "Algorithm," Merriam-Webster.com, accessed November 4, 2022, https://www.merriam-webster.com/dictionary/algorithm.
30. Jason Brownlee, "Difference Between Algorithm and Model in Machine Learning," *Machine Learning Mastery* (blog), April 28, 2020, https://machinelearningmastery.com/difference-between-algorithm-and-model-in-machine-learning.
31. Rayna Hollander, "Amazon Is Improving the Accuracy of Alexa's Natural Language Understanding," *Business Insider*, October 11, 2019, https://www.businessinsider.com/amazon-bolsters-alexa-skill-voice-accuracy-2019-10.
32. Lee Rainie et al., "AI and Human Enhancement: Americans' Openness Is Tempered by a Range of Concerns," Pew Research Center, March 17, 2022, https://www.pewresearch.org/internet/2022/03/17/ai-and-human-enhancement-americans-openness-is-tempered-by-a-range-of-concerns.
33. Melissa Bauman, "Why Waiting for Perfect Autonomous Vehicles May Cost Lives," *The RAND Blog*, November 7, 2017, https://www.rand.org/pubs/articles/2017/why-waiting-for-perfect-autonomous-vehicles-may-cost-lives.html.
34. "Practices of Science: False Positives and False Negatives," University of Hawai'i at Manoa, accessed November 3, 2022, https://manoa.hawaii.edu/exploringourfluidearth/chemical

/matter/properties-matter/practices-science-false-positives-and-false-negatives.
35. Noam Bressler and Shlomo Tanor, "A Guide to Evaluation Metrics for Classification Models," *Deep Checks*, April 14, 2021, https://deepchecks.com/a-guide-to-evaluation-metrics-for-classification-models.
36. Bressler and Tanor, "A Guide to Evaluation Metrics."
37. Zeya LT, "Essential Things You Need to Know About F1-Score," *Medium*, November 23, 2021, https://towardsdatascience.com/essential-things-you-need-to-know-about-f1-score-dbd973bf1a3. As mentioned, there are many metrics beyond the illustrative examples listed here. Interested policymakers can look into further evaluation metrics, including area under the curve, receiver operating characteristic curve, mean squared error, mean absolute error, and confusion matrices, among other useful metrics. Policymakers should consider their purposes, the needs of certain applications, and which metrics are best suited for those needs.
38. Ali Borji, "Pros and Cons of GAN Evaluation Measures," *Computer Vision and Image Understanding* 179 (February 2019): 41–65, https://www.sciencedirect.com/science/article/abs/pii/S1077314218304272.
39. Inioluwa Deborah Raji et al., "AI and the Everything in the Whole Wide World Benchmark," *OpenReview.net*, modified October 21, 2021, https://openreview.net/forum?id=j6NxpQbREA1.
40. "ImageNet Large Scale Visual Recognition Challenge," *ImageNet*, accessed November 8, 2022, https://www.image-net.org/challenges/LSVRC/index.php.
41. Allen Institute for AI, "HellaSwag," accessed May 20, 2024, https://allenai.org/data/hellaswag.
42. Ben Dickson, "Why We Must Rethink AI Benchmarks," *TechTalks* (blog), December 6, 2021, https://bdtechtalks.com/2021/12/06/ai-benchmarks-limitations.
43. Raji et al., "AI and the Everything in the Whole Wide World Benchmark."
44. National Artificial Intelligence Initiative, "About Artificial Intelligence," accessed November 1, 2022, https://www.ai.gov/about.
45. National Institute of Standards and Technology, "AI Risk Management Framework," accessed May 20, 2024, https://www.nist.gov/itl/ai-risk-management-framework.
46. The White House, "Blueprint for an AI Bill of Rights—OSTP," accessed November 1, 2022, https://www.whitehouse.gov/ostp/ai-bill-of-rights.
47. Jamie Baker, "A DPA for the 21st Century," Center for Security and Emerging Technology, April 2021, https://cset.georgetown.edu/publication/a-dpa-for-the-21st-century/.
48. Federal Communications Commission, "FCC Confirms That TCPA Applies to AI Technologies That Generate Human Voices," Declaratory Ruling, accessed May 20, 2024, https://www.fcc.gov/document/fcc-confirms-tcpa-applies-ai-technologies-generate-human-voices.
49. Matthew Feeney, "Deepfake Laws Risk Creating More Problems Than They Solve," Regulatory Transparency Project, March 1, 2021, https://www.cato.org/sites/cato.org/files/2021-03/Paper-Deepfake-Laws-Risk-Creating-More-Problems-Than-They-Solve.pdf.
50. National Conference of State Legislatures, "Artificial Intelligence (AI) in Elections and Campaigns," updated April 30, 2024, https://www.ncsl.org/elections-and-campaigns/artificial-intelligence-ai-in-elections-and-campaigns.
51. National Conference of State Legislatures, "Autonomous Vehicles—Self-Driving Vehicles Enacted Legislation," updated February 18, 2020, https://www.ncsl.org/research/transportation/autonomous-vehicles-self-driving-vehicles-enacted-legislation.aspx.
52. J. Edward Moreno, "New York City AI Bias Law Charts New Territory for Employers," Daily Labor Report, *Bloomberg Law*, August 29, 2022, https://news.bloomberglaw.com/daily

-labor-report/new-york-city-ai-bias-law-charts-new-territory-for-employers.

53. Pub. L. No. 117–167 (2022), making appropriations for the legislative branch for the fiscal year ending September 30, 2022, and for other purposes.
54. Melanie Lefkowitz, "Professor's Perceptron Paved the Way for AI—60 Years Too Soon," *Cornell Chronicle*, September 25, 2019, https://news.cornell.edu/stories/2019/09/professors-perceptron-paved-way-ai-60-years-too-soon.
55. Defense Advanced Research Projects Agency, "The Grand Challenge," accessed November 1, 2022, https://www.darpa.mil/about-us/timeline/-grand-challenge-for-autonomous-vehicles.
56. Ben Leonard and Ruth Reader, "Artificial Intelligence Was Supposed to Transform Health Care. It Hasn't," *Politico*, August 15, 2022, https://www.politico.com/news/2022/08/15/artificial-intelligence-health-care-00051828.
57. Megan Lewis, "Why It's a Problem That Pulse Oximeters Don't Work as Well on Patients of Color," *MIT News*, August 2, 2022, https://news.mit.edu/2022/pulse-oximeters-dont-work-as-well-patients-of-color-0802.
58. Jeffrey Dastin, "Amazon Scraps Secret AI Recruiting Tool That Showed Bias against Women," *Reuters*, October 10, 2018, https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G.
59. Pádraig Belton, "The Computer Chip Industry Has a Dirty Climate Secret," *The Guardian*, September 18, 2021, https://www.theguardian.com/environment/2021/sep/18/semiconductor-silicon-chips-carbon-footprint-climate.
60. American Trucking Associations, "Economics and Industry Data," accessed November 1, 2022, https://www.trucking.org/economics-and-industry-data.
61. Joe Hernandez, "A Military Drone with a Mind of Its Own Was Used in Combat, U.N. Says," *NPR*, June 1, 2021, https://www.npr.org/2021/06/01/1002196245/a-u-n-report-suggests-libya-saw-the-first-battlefield-killing-by-an-autonomous-d.
62. ARM, "What Is AI Inference?," accessed November 30, 2022, https://www.arm.com/glossary/ai-inference.
63. Scott Cleland, "Google's "Infringenovation" Secrets," *Forbes*, October 4, 2011, https://www.forbes.com/sites/scottcleland/2011/10/03/googles-infringenovation-secrets/.
64. Rebecca Laborde, "The Three V's of Big Data: Volume, Velocity, and Variety," *Oracle Life Sciences Blog*, January 23, 2020, https://blogs.oracle.com/life-sciences/post/the-three-vx27s-of-big-data-volume-velocity-and-variety. The three V's have steadily expanded over time, with some sources adding qualities such as veracity and value. Both are important qualities to be sure, but this study uses the original three for the sake of simplicity.
65. Rony Chow, "ImageNet: A Pioneering Vision for Computers," *History of Data Science*, August 27, 2021, https://www.historyofdatascience.com/imagenet-a-pioneering-vision-for-computers.
66. "From Not Working to Neural Networking," *The Economist*, June 23, 2016, https://www.economist.com/special-report/2016/06/23/from-not-working-to-neural-networking.
67. "How Much Training Data Is Required for Machine Learning Algorithms?," *Cogito Tech* (blog), July 9, 2019, https://www.cogitotech.com/blog/how-much-training-data-is-required-for-machine-learning-algorithms.
68. Sébastien Bubeck and Mark Sellke, "A Universal Law of Robustness via Isoperimetry," Microsoft Research, December 1, 2021, https://www.microsoft.com/en-us/research/publication/a-universal-law-of-robustness-via-isoperimetry.
69. Bubeck and Sellke, "A Universal Law."
70. Bubeck and Sellke, "A Universal Law."
71. There are many rules of thumb that have developed to estimate the data needed; for instance, one source recommends as much data as 10 times the number of parameters required by the model. Such a recommendation is not

based in science, though that does not mean it is not a useful goal.

72. A classic debate in modern AI is the tradeoff between big data AI innovation and personal privacy. Some have argued that small data strategies offer a potential future for AI that preserves both innovation and privacy. Others contend that small data might allow democracies to compete against unrestricted authoritarian data collection practices without sacrificing democratic principles. The ultimate impact and viability of these strategies, however, remains to be seen.
73. The idea of using artificial data to train systems may seem odd at first. Real-world data, however, are not always needed to learn the required lesson or skill. The textbooks used in school, for instance, often teach using artificial data—an economics problem in a textbook rarely uses real-world data, but the principles of that problem can be instructive, nonetheless.
74. Husanjot Chahal, Helen Toner, and Illya Rahkovsky, "Small Data's Big AI Potential," *Center for Security and Emerging Technology*, September 2022, https://cset.georgetown.edu/publication/small-datas-big-ai-potential.
75. Stanford University Human-Centered Artificial Intelligence, "Artificial Intelligence Index Report 2024," accessed May 20, 2024, https://aiindex.stanford.edu/wp-content/uploads/2024/04/HAI_AI-Index-Report-2024.pdf.
76. David Silver et al., "AlphaZero: Shedding New Light on Chess, Shogi, and Go," *DeepMind* (blog), December 6, 2018, https://www.deepmind.com/blog/alphazero-shedding-new-light-on-chess-shogi-and-go.
77. IBM, "What Is Zero-Shot Learning?," January 24, 2024, https://www.ibm.com/topics/zero-shot-learning.
78. IBM, "What Is Few-Shot Learning?," accessed January 24, 2024, https://www.ibm.com/topics/few-shot-learning#:~:text=Few%2Dshot%20learning%20is%20a,suitable%20training%20data%20is%20scarce.
79. Patrick Grother, Mei Ngan, and Kayee Hanaoka, "Face Recognition Vendor Test (FRVT) Part 3: Demographic Effects," NSTIR 8280, National Institute of Standards and Technology, December 2019, https://nvlpubs.nist.gov/nistpubs/ir/2019/nist.ir.8280.pdf.
80. Jamie Baker, Laurie Hobart, and Matthew Mittelsteadt, "AI for Judges," Center for Security and Emerging Technology, December 2021, https://cset.georgetown.edu/publication/ai-for-judges.
81. "What Is Data Velocity?: Data Defined," *Indicative*, accessed November 29, 2022, https://www.indicative.com/resource/data-velocity.
82. Sophia Y. Wang, Suzann Pershing, and Aaron Y. Lee, "Big Data Requirements for Artificial Intelligence," *Current Opinion in Ophthalmology* 31, no. 5 (2020): 318–23.
83. Jamie Baker, "Symposium Report: National Security Law and the Coming AI Revolution," Center for Security and Emerging Technology, April 13, 2021, https://cset.georgetown.edu/article/symposium-report-national-security-law-and-the-coming-ai-revolution.
84. Nathaniel Kim, Insrup Lee, and Javier Zazo, "Technology Factsheet: Internet of Things," Belfer Center for Science and International Affairs, June 2019, https://www.belfercenter.org/publication/technology-factsheet-internet-things.
85. "What Is a Data Warehouse?," *Oracle*, accessed November 29, 2022, https://www.oracle.com/database/what-is-a-data-warehouse.
86. The infrastructure that makes AI possible is not immune to unintended social effects. Data warehouses have been noted for their constant hum, which disrupts both wildlife and local residents. Warehousing also creates electronic waste and, when using fissile fuels, yields a high carbon footprint. With large data and computation demands, AI systems may create many externalities in the communities that support their infrastructural operation.
87. Husanjot Chahal, Ryan Fedasiuk, and Carrick Flynn, "Messier Than Oil: Assessing Data

Advantage in Military AI," Center for Security and Emerging Technology, July 2020, https://cset.georgetown.edu/publication/messier-than-oil-assessing-data-advantage-in-military-ai.

88. This tedious task is often assigned to research assistants or Amazon's Mechanical Turk on-demand crowdsourcing service.
89. Satyam Kumar, "7 Ways to Handle Missing Values in Machine Learning," *Towards Data Science* (blog), July 24, 2020.
90. This is a more difficult task than it would appear. Many engineers would label a yellow banana as just "banana," while labeling an unripe banana "green banana." In either case, "banana" is technically appropriate, but convention leads us to qualify the unripe version with the adjective "green" while leaving the ripe version unqualified. Labeling is a deeply human task informed by the viewpoint and habits of the engineer and the culture of the engineer.
91. OpenAI, "DALL·E 3," accessed May 22, 2024, https://openai.com/index/dall-e-3.
92. Jake Silberg and James Manyika, "Tackling Bias in Artificial Intelligence (and in Humans)," McKinsey Global Institute, McKinsey & Company, June 6, 2019, https://www.mckinsey.com/featured-insights/artificial-intelligence/tackling-bias-in-artificial-intelligence-and-in-humans.
93. Baker, Hobart, and Mittelsteadt, "AI for Judges."
94. Reva Schwartz et al., "Towards a Standard for Identifying and Managing Bias in Artificial Intelligence," NIST Special Publication, National Institute of Standards and Technology, 2022, https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1270.pdf.
95. Nicole Turner Lee, Paul Resnick, and Genie Barton, "Algorithmic Bias Detection and Mitigation: Best Practices and Policies to Reduce Consumer Harms," Brookings Institution, May 22, 2019, https://www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms.
96. Correcting bias can itself be marred by bias. A given application will have an unknown number of unknown biases that need correcting. Only the biases of which one is aware will gain attention. Further, "fairness" of results can be difficult to define and often requires subjective decision-making that itself is naturally biased.
97. Elham Tabassi et al., "A Taxonomy and Terminology of Adversarial Machine Learning," National Institute of Standards and Technology, October 2019, https://www.researchgate.net/publication/344943809_A_taxonomy_and_terminology_of_adversarial_machine_learning. All algorithms are vulnerable to traditional cyberattacks. AI systems are no different. While data poisoning attacks are often pointed to as the marquee AI vulnerability, policymakers must not forget that more traditional exploits remain. Data-based attacks represent an added layer of insecurity unique to AI systems on top of the range of older, still relevant cyber vulnerabilities. Also, recall that AI algorithms depend on other systems. An attacker that cannot gain access to an AI or its training data can still attack the system by compromising connected devices.
98. Shaun Waterman, "Hacking Poses Risks for Artificial Intelligence," Center for Security and Emerging Technology, March 1, 2022, https://cset.georgetown.edu/article/hacking-poses-risks-for-artificial-intelligence.
99. Tim G. J. Rudner and Helen Toner, "Key Concepts in AI Safety: Robustness and Adversarial Examples," Center for Security and Emerging Technology, March 1, 2021, https://cset.georgetown.edu/publication/key-concepts-in-ai-safety-robustness-and-adversarial-examples.
100. Kevin Eykholt et al., "Robust Physical-World Attacks on Deep Learning Visual Classification," April 10, 2018, https://doi.org/10.48550/arXiv.1707.08945.

101. Glaze, "What Is Glaze?," accessed May 21, 2024, https://glaze.cs.uchicago.edu/what-is-glaze.html.
102. Federal Trade Commission, "The FTC Voice Cloning Challenge," accessed May 21, 2024, https://www.ftc.gov/news-events/contests/ftc-voice-cloning-challenge.
103. In many cases, and across industries, governments and government-commissioned industry regulators control standards. Standardization is deeply entwined with policy. Most standards were decided without considering the needs of AI, yet their impact on AI could be great. Each field and potential AI application may benefit from looking at how their designs and choices might affect these systems.
104. Jason Fernando, "GAAP: Understanding It and the 10 Key Principles," *Investopedia*, updated June 28, 2022, https://www.investopedia.com/terms/g/gaap.asp.
105. "LIFO vs. FIFO," Corporate Finance Institute, updated January 6, 2023, https://corporatefinanceinstitute.com/resources/accounting/lifo-vs-fifo.
106. Charlene Rhinehart, "Does US GAAP Prefer FIFO or LIFO Accounting?," *Investopedia*, updated July 31, 2021, https://www.investopedia.com/ask/answers/032415/does-us-gaap-prefer-fifo-or-lifo-accounting.asp.
107. Yong-Yeon Jo et al., "Impact of Image Compression on Deep Learning-Based Mammogram Classification," *Scientific Reports* 11, no. 1 (2021): 7924.
108. Tom Simonite, "Apple's Latest iPhones Are Packed with AI Smarts," *Wired*, September 12, 2018, https://www.wired.com/story/apples-latest-iphones-packed-with-ai-smarts.
109. Not all AI uses machine learning. Deep Blue is an example of an expert system, a form of "symbolic" AI that does not use machine learning. Expert systems are designed with a vast knowledge base that, ideally, can be relied on in most cases. When faced with uncertainty, these systems turn to an "inference engine," which infers the best action on the basis of existing knowledge and guidance from a set of if-then rules. The term "inference," still in use today, is derived from this function. Deep Blue's chess chip was specifically optimized to efficiently search this massive knowledge base and execute its inference engine rules. While the expert system approach is largely outdated, many of its techniques have been adopted by varieties of machine learning in use today.
110. *Encyclopedia Britannica*, 2022, s.v. "Semiconductor," https://www.britannica.com/science/semiconductor.
111. *Encyclopedia Britannica*, 2022, s.v. "Moore's law," https://www.britannica.com/technology/Moores-law.
112. Saif M. Kahn, "AI Chips: What They Are and Why They Matter," Center for Security and Emerging Technologies, April 2020, https://cset.georgetown.edu/publication/ai-chips-what-they-are-and-why-they-matter.
113. Jennifer Prendki, "Why the End of Moore's Law Means the End of Big Data as We Know It," *Alectio*, March 27, 2020, https://alectio.com/2020/03/27/why-the-end-of-moores-law-means-the-end-of-big-data-as-we-know-it.
114. Methods used to improve semiconductor speeds beyond merely shrinking transistors can add complexity to chips, naturally decreasing their security. As a rule of thumb, simplicity drives security. In recent years, many chips have implemented a speed-boosting technique called "speculative execution." The complexity of this technique opened the door to a variety of hardware-based cyberattacks called "transient execution CPU vulnerabilities." The most famous examples are Spectre and Meltdown, which are thus far incurable classes of vulnerabilities that expose nearly every device running the iOS, Linux, macOS, and Windows operating systems. Hardware-based vulnerabilities such as these are difficult to mitigate because potential solutions may require physically altering the devices. With the explosion of chips in use in increasingly diverse AI systems, such vulnerabilities can be costly to mitigate.

115. Andrew Lohn and Micah Musser, "AI and Compute: How Much Longer Can Computing Power Drive Artificial Intelligence Progress?," Center for Security and Emerging Technology, January 2022., https://cset.georgetown.edu/publication/ai-and-compute.
116. Brian Bailey, "Von Neumann Is Struggling," *Semiconductor Engineering* (blog), January 18, 2021, https://semiengineering.com/von-neumann-is-struggling.
117. Tom Simonite, "An Old Technique Could Put Artificial Intelligence in Your Hearing Aid," *Wired*, November 27, 2017, https://www.wired.com/story/an-old-technique-could-put-artificial-intelligence-in-your-hearing-aid.
118. In recent years, AI systems have sometimes been criticized for their high energy toll and resulting carbon footprint. In general, the energy used by a system is directly related to its speed and efficiency. The faster a chip completes its computations, the less energy it uses and, by extension, the smaller its carbon footprint. Complicating this picture is algorithmic efficiency. A chip can be fast, but if the algorithms it runs are slow, any chip-side speed improvements might not decrease total energy use. Emerging designs such as analog chips could purportedly slash the energy used by chips; however, these designs have yet to affect the mainstream.
119. GPUs are sought after by a wide variety of interest groups and are often in short supply. GPUs' special functions are widely used in computationally intensive fields, including video gaming and cryptocurrency mining. As a result, crypto miners and gamers can crowd out supply that might otherwise be used for AI. Because semiconductor fabrication cannot be easily ramped up and down to meet competing demands, the supply is tightly limited. When supply is tight, chips provided to one application sector directly crowd out chips provided to another.
120. It is very common for data to be arranged, and by extension analyzed, as a matrix of numbers. An example illustration can be found in the "Data" section of this report.
121. Falan Yinug, "Semiconductors: A Strategic U.S. Advantage in the Global Artificial Intelligence Technology Race," white paper, Semiconductor Industry Association, August 2018, https://www.semiconductors.org/wp-content/uploads/2018/08/81018_SIA_AI_white_paper_-_FINAL_08092018_with_all_member_edits_with_logo3.pdf.
122. Mike Brogioli, "The DSP Hardware/Software Continuum," in *DSP for Embedded and Real-Time Systems*, ed. Robert Oshana (Oxford, UK: Newnes, 2012), 103–12.
123. "ASIC vs. FPGA: What's the Difference?," *AsicNorth*, October 10, 2020, https://www.asicnorth.com/blog/asic-vs-fpga-difference.
124. Gaurav Batra et al., "AI Hardware: Value Creation for Semiconductor Companies," McKinsey & Company, January 2, 2019, https://www.mckinsey.com/industries/semiconductors/our-insights/artificial-intelligence-hardware-new-opportunities-for-semiconductor-companies.
125. Batra et al., "AI Hardware."
126. "Introduction to Semiconductors," *AMD*, accessed December 31, 2021, https://www.amd.com/en/technologies/introduction-to-semiconductors.
127. *Encyclopedia Britannica*, 2020, s.v. "Dopant," https://www.britannica.com/technology/dopant.
128. "Understanding RAM and DRAM Computer Memory Types," *ATP*, July 1, 2022, https://www.atpinc.com/blog/computer-memory-types-dram-ram-module.
129. A standard CPU includes many important units. These include multiple cores, each representing a complete processing unit, allowing the CPU to run multiple instructions and programs at once. Clock generator chips keep time and set the pace of computation. Address generation units calculate where information is stored in memory. Interconnects are the wires that ferry data and tie all these components together.

130. Will Hunt and Remco Zwetsloot, "The Chipmakers: U.S. Strengths and Priorities for the High-End Semiconductor Workforce," Center for Security and Emerging Technology, September 2020, https://cset.georgetown.edu/publication/the-chipmakers-u-s-strengths-and-priorities-for-the-high-end-semiconductor-workforce.
131. Hunt and Zwetsloot, "The Chipmakers."
132. Kahn, "AI Chips"; Prendki, "Why the End of Moore's Law Means the End of Big Data as We Know It."
133. "Semiconductor Fabrication: How Are They Manufactured?," *Halocarbon*, December 12, 2022, https://halocarbon.com/semiconductor-fabrication-how-are-they-manufactured.
134. Khan, "AI Chips."
135. David Rotman, "We're Not Prepared for the End of Moore's Law," *MIT Technology Review*, February 24, 2020, https://www.technologyreview.com/2020/02/24/905789/were-not-prepared-for-the-end-of-moores-law/.
136. Hunt and Zwetsloot, "The Chipmakers."
137. Will Hunt, "Sustaining U.S. Competitiveness in Semiconductor Manufacturing," Center for Security and Emerging Technologies, January 2022, https://cset.georgetown.edu/publication/sustaining-u-s-competitiveness-in-semiconductor-manufacturing.
138. Katie Tarasov, "ASML Is the Only Company Making the $200 Million Machines Needed to Print Every Advanced Microchip. Here's an Inside Look," *CNBC*, March 23, 2022, https://www.cnbc.com/2022/03/23/inside-asml-the-company-advanced-chipmakers-use-for-euv-lithography.html.
139. Semiconductor supply chains are long, complex, and brittle. Throughout the chain, malicious actors can inject vulnerabilities directly into chips. Given the complexity of supply chains, alterations to chips can be difficult to spot, potentially allowing insecurities to enter systems unnoticed and persist for years. In 2018, Bloomberg released a controversial report claiming that supermicro-manufactured semiconductors had been implanted for years with a microscopic chip that could instruct a system to communicate with certain external servers without user consent. The report implicated the Chinese government as the perpetrator. While the intelligence community disputes this report, it is illustrative of the vulnerability of chip supply chains and the persistent vulnerabilities that supply chain attacks can create.
140. Andrew Lohn and Micah Musser, "AI and Compute: How Much Longer Can Computing Power Drive Artificial Intelligence Progress?," Center for Security and Emerging Technology, January 2022, https://cset.georgetown.edu/publication/ai-and-compute.
141. Kim Martineau, "What Is Federated Learning?," *IBM Research* (blog), August 24, 2022, https://research.ibm.com/blog/what-is-federated-learning.
142. Marta Garnelo and Murray Shanahan, "Reconciling Deep Learning with Symbolic Artificial Intelligence: Representing Objects and Relations," *Current Opinion in Behavioral Sciences* 29 (October 2019): 17–23, https://www.sciencedirect.com/science/article/pii/S2352154618301943?via%3Dihub.
143. scikit-learn, "Clustering," accessed December 14, 2022, https://scikit-learn.org/stable/modules/clustering.html.
144. The features discovered by these networks will rarely be defined in human-familiar terms. AI systems fundamentally see the world differently. Where we might see a blue eye, an AI system might instead see a circle of pixels and the numerical values that represent the pixel colors. Nonetheless, it can still learn this pattern and understand the conclusions it correlates with. Present-day AI systems cannot, however, say *why* a pattern is meaningful. AI can reach correct conclusions without any true understanding.
145. The difference between the terms artificial intelligence (AI), machine learning (ML), and artificial neural networks (ANNs) can seem fuzzy and confusing. It is often helpful to think hierarchically. At the top is AI, which

is the overriding goal of all of these technologies. Next is ML, the general-purpose technology used to achieve this goal of AI. Beneath ML is supervised, reinforcement, and unsupervised learning, all of which are distinguished by their strategies for achieving intelligent outcomes. Beneath these, one can find more specific models. Deep learning can be thought of as a subcategory of any of the previous three approaches to machine learning. Reinforcement learning can use deep learning, as can supervised learning. Deep learning is distinguished not by its approach but by its tools—specifically, its use of ANNs. As ML and deep learning have become exceedingly popular in AI, these terms have become increasingly interchangeable in casual use. Most people mix these terms, and few—especially in policy settings—will bat an eye at the inevitable confusion.

146. IBM, "What Is Deep Learning?," accessed February 22, 2021, https://www.ibm.com/cloud/learn/deep-learning.
147. Karen Hao, "This Is How AI Bias Really Happens—and Why It's So Hard to Fix," *MIT Technology Review*, February 4, 2019, https://www.technologyreview.com/2019/02/04/137602/this-is-how-ai-bias-really-happensand-why-its-so-hard-to-fix.
148. Adam Zewe, "Can Machine-Learning Models Overcome Biased Datasets?," *MIT News*, February 21, 2022, https://news.mit.edu/2022/machine-learning-biased-data-0221.
149. Nicol Turner Lee, Paul Resnick, and Genie Barton, "Algorithmic Bias Detection and Mitigation: Best Practices and Policies to Reduce Consumer Harms," Brookings Institution, May 22, 2019, https://www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms.
150. Gayane Grigoryan and Andrew J. Collins, "Is Explainability Always Necessary? Discussion on Explainable AI," Old Dominion University, April 14, 2022, https://digitalcommons.odu.edu/cgi/viewcontent.cgi?article=1013&context=msvcapstone.
151. Hima Lakkaraju, "Stanford Seminar—ML Explainability Part 1—Overview and Motivation for Explainability," Stanford University, 2022, https://www.youtube.com/watch?v=_DYQdP_F-LA.
152. scikit-learn, "Decision Trees," accessed December 12, 2022, https://scikit-learn.org/stable/modules/tree.html.
153. "Algorithm Descriptions," *Captum*, accessed December 11, 2022, https://captum.ai.
154. Narine Kokhlikyan and Ludwig Schubert, "Opening Up the Black Box: Model Understanding with Captum and PyTorch," presentation at GTC 2020, *PyTorch* YouTube channel, July 1, 2020, https://www.youtube.com/watch?v=0QLrRyLndFI.
155. Carlos Ignacio Gutierrez and Gary E. Marchant, "A Global Perspective of Soft Law Programs for the Governance of Artificial Intelligence," Sandra Day O'Connor College of Law, Arizona State University, May 28, 2021, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3855171.
156. Michèle A. Flournoy, Avril Haines, and Gabrielle Chefitz, "Building Trust through Testing," Center for Security and Emerging Technologies, October 2020, https://cset.georgetown.edu/wp-content/uploads/Building-Trust-Through-Testing.pdf.
157. Alex Engler, "Auditing Employment Algorithms for Discrimination," Brookings Institution, March 12, 2021, https://www.brookings.edu/research/auditing-employment-algorithms-for-discrimination.
158. Schwartz et al., "Towards a Standard for Identifying and Managing Bias."
159. Marietje Schaake and Jack Clark, "Stanford Launches AI Audit Challenge," Stanford University Human-Centered Artificial Intelligence, July 11, 2022, https://hai.stanford.edu/news/stanford-launches-ai-audit-challenge.
160. In a realistic sense, the activation function isn't a trigger but the mathematical rules of the road

for the type of output the input data is mathematically transformed into. In this case, one set of conditions will lead the function to transform the input into a prime/subprime lending decision. Here the function simply produces a **binary** either/or output. The world isn't always so black and white, however. In other cases, an activation function might allow the data to be transformed into one of many choices—its value representing the probability that the pattern represents a certain prediction. Activation functions can take many forms. In sum, the activation is more or less a prediction-formatting mechanism. The type of prediction wanted is determined by this function.

161. In a more mathematical sense, the bias also serves to orient the function toward reality. Beneath each function is a graph, and the direction and starting values of the graph are set values using an intercept that is this bias value.
162. IBM, "What Is Supervised Learning?," 2020, https://www.ibm.com/cloud/learn/supervised-learning.
163. IBM, "What Is Supervised Learning?
164. "Huge 'Foundation Models' Are Turbo-Charging AI Progress," *The Economist*, June 11, 2022, https://www.economist.com/interactive/briefing/2022/06/11/huge-foundation-models-are-turbo-charging-ai-progress.
165. Separate from the training and testing data are **validation data**, which are used by the engineer after the training process to adjust and tune the hyperparameters. Only after training and validation is the unused test set used to provide a final measure of the model.
166. scikit-learn, "Machine Learning in Python—scikit-learn 1.2.0 Documentation," 2022, https://scikit-learn.org/stable/index.html.
167. Ben Dickson, "What's the Transformer Machine Learning Model? And Why Should You Care?," *TNW*, May 3, 2022, https://thenextweb.com/news/whats-the-transformer-machine-learning-model.
168. Jason Brownlee, "A Gentle Introduction to Generative Adversarial Networks (GANs)," *Machine Learning Mastery* (blog), updated July 19, 2019, https://machinelearningmastery.com/what-are-generative-adversarial-networks-gans.
169. Sumit Saha, "A Comprehensive Guide to Convolutional Neural Networks—the ELI5 Way," *Medium*, accessed November 16, 2022, https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53.
170. Simeon Kostadinov, "How Recurrent Neural Networks Work," *Medium*, accessed November 10, 2019, https://towardsdatascience.com/learn-how-recurrent-neural-networks-work-84e975feaaf7.
171. Ben Dickson, "What Is Semi-Supervised Machine Learning?," *TechTalks*, January 4, 2021, https://bdtechtalks.com/2021/01/04/semi-supervised-machine-learning.
172. National Artificial Intelligence Initiative Act of 2020, Pub. L. No. H.R. 6216 (2020).
173. Google Developer Program, "Machine Learning: Foundational Courses, Generalization," accessed May 20, 2024, https://developers.google.com/machine-learning/crash-course/generalization/video-lecture#:~:text=Generalization%20refers%20to%20your%20model's,Time%3A%205%20minutes%20Learning%20Objectives.
174. Pethokoukis, "How AI Is Like That Other General Purpose Technology, Electricity"; Elhanan Helpman, ed., *General Purpose Technologies and Economic Growth* (Cambridge, MA: MIT Press, 2003), https://www.aei.org/economics/how-ai-is-like-that-other-general-purpose-technology-electricity.
175. Jason Brownlee, "Difference between Algorithm and Model in Machine Learning," *Machine Learning Mastery* (blog), April 28, 2020, https://machinelearningmastery.com/difference-between-algorithm-and-model-in-machine-learning.

About the Author

Matthew Mittelsteadt is a technologist and research fellow at the Mercatus Center at George Mason University whose work focuses on artificial intelligence, cybersecurity, and technology policy. Prior to joining Mercatus, Matthew worked as a fellow at the Institute for Security, Policy, and Law, where he studied the use of AI in the courtroom and the technical implementation of AI arms control. His work has appeared in *The Hill*, *American Banker*, and the *New York Daily News*. In the private sector, he has worked as an information technology project manager. He holds an MS in cybersecurity from New York University, an MPA from Syracuse University, and a BA in economics and Russian studies from St. Olaf College.

About the Mercatus Center at George Mason University

The Mercatus Center at George Mason University is the world's premier university source for market-oriented ideas—bridging the gap between academic ideas and real-world problems.

A university-based research center, the Mercatus Center advances knowledge about how markets work to improve people's lives by training graduate students, conducting research, and applying economics to offer solutions to society's most pressing problems.

Our mission is to generate knowledge and understanding of the institutions that affect the freedom to prosper, and to find sustainable solutions that overcome the barriers preventing individuals from living free, prosperous, and peaceful lives.

Founded in 1980, the Mercatus Center is located on George Mason University's Arlington and Fairfax campuses.