MERCATUS SPECIAL STUDY

ARTIFICIAL INTELLIGENCE

AN INTRODUCTION FOR POLICYMAKERS

Matthew Mittelsteadt, ***Mercatus Center***

MERCATUS.ORG

MERCATUS CENTER
George Mason University

Matthew Mittelsteadt, "Artificial Intelligence: An Introduction for Policymakers," Mercatus Special Study, Mercatus Center at George Mason University, Arlington, VA, February 2023.

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

Over the past decade, real-world artificial intelligence (AI) has evoked electronic assistants—such as Alexa from Amazon or Siri from Apple—in the public's mind. But today, society is at an inflection point. In 2021, Stanford University's Human-Centered Artificial Intelligence Institute wrote about an AI "paradigm shift."[1] Its report identified the rise of what it termed **foundation models**, large-scale systems trained on broad sets of data that can be easily adapted to a wide range of downstream tasks.[2] Armed with computational heft and flexibility, this new class of models could offer many of the tools needed for AI to step beyond a mere curiosity. Even if one reserves a measure of skepticism toward AI hype, many applications that debuted in 2022 promise new possibilities in several domains of application:

- Midjourney's art generator produced near-human-quality works.
- AlphaFold predicts the structure of nearly every known protein, a critical new tool in biological and medical research.
- AlphaTensor discovered a more efficient approach to matrix multiplication than previously known, and this could soon speed up a wide range of applications.
- MūtCompute discovered an enzyme that breaks down polyethylene terephthalate, a common plastic that represents 12 percent of global waste.
- AlphaCode ranked within the top 54 percent of participants in competitive coding competitions with its increasingly efficient self-generated algorithms.

- Codex translates natural language into code, potentially opening engineering to a wider audience.
- OpenAI's ChatGPT produces medium-length, logically complete responses to complex text prompts.

The breadth of fields of applications is worth noting. 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."[3,4] 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.

Critics claim that these advances in AI are skin deep, mere "**stochastic parrots**" that randomly rearrange and regurgitate data.[5] They may look effective, critics argue, but lack any true understanding, common sense, or ability to explain their decisions. The critics could very well be correct about artificial intelligence lacking intelligence, but the critics will err dramatically if they dismiss AI outright as unimportant.

Yet, policymakers are not keeping up with all these developments. Knowledge is necessary but not sufficient for good governance. Even if lawmakers were to grasp basic notions of AI engineering and acquire a sense of the depth and breadth of AI's effect, it is an open question whether they would be able to translate that knowledge into a consensus for AI governance. After all, Silicon Valley, as a collective of entrepreneurs and innovators who better understand the mechanics and effects of AI, has not arrived at a consensus on the governance of AI either. In this work, we hope to impart some basic knowledge on AI design, application, and policy challenges to inform policy-minded readers. We do this without naivete, because we are deeply aware that the politics of policy design could be a problem more complex than an understanding of some of the most sophisticated AI systems.

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."[6]
>
> —Jacob Helberg, former Google news policy lead; commissioner, US China Commission

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

In summer 2022, AI art 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 with art so generated, while comforting others with the understanding that generative AI was still out of immediate reach. Yet, in just a matter of weeks, things changed. As OpenAI broadened access to the full version of DALL·E 2, Midjourney's beta app generated covers for *The Economist*,[7] and Hugging Face released the powerful Stable Diffusion, these wonky generators suddenly proved capable. Often, their outputs were professional quality and, in one instance, even "skilled" enough to

win a state art competition.[8] The progress of this technology moved at an astounding pace.

This sudden burst of innovation likely took those working in arts policy off guard. In a matter of weeks, policymakers had to shift gears toward confronting a slew of novel AI-based issues that they perhaps wouldn't have considered just months earlier. One such controversy is artistic rights. It was found that the engineers had built these systems by "training" the AI to produce art based on preexisting human-crafted works scraped from the web. Often, this process was undertaken without artistic consent. As a result, prominent digital artists found this software could produce near-perfect renditions of their works, allowing anyone to appropriate their signature styles if in possession of the necessary know-how and the computational capacity.[9] This situation raised questions of usage rights, privacy, personal autonomy, and copyright infringement.

Many affected artists view this situation as potentially existential. To those working at the top levels of AI policy, it remains off-radar. When interviewed on the effect of AI art generators, 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.[10] One is tempted to believe that such an important question was never discussed by the broader task force.

The reason? AI has been treated as a specialty. Because the task force is composed almost exclusively of computer scientists, one is hardly surprised that it was not thinking about artistic rights questions. Had AI been viewed as having general-purpose effects, perhaps those in the arts would have been engaged and their voices heard in the design of solutions to those problems. Breadth of expertise, however, cannot sacrifice depth of technical knowledge. Only by understanding the scientific progress of art generators—how data are scraped and used to train AI, what type of data is needed—could those concerned about artistic rights have predicted this issue and have begun to consider appropriate action. Many of these art generators have now been open sourced, meaning their code is no longer controlled by a single entity, and affected artists may have little recourse. Appropriate policy would have required engaging the specific art-generator application. Specificity is currently missing from AI policy design.

THE IMPORTANCE OF DEEPER UNDERSTANDING

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."[11] All policy areas will be touched and even transformed by artificial intelligence (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 the computer scientists have done so only at a high level, a so-called Level 1 understanding. They can engage with the concept, and perhaps entertain abstract effects, but cannot dig into problems nor imagine specific solutions to them. AI is maturing, and policymakers should go deeper. The goal is 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 an AI chat bot, they know how one functions. Although they have not

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 its 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 Develpment is researching 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 researches the possible effects of widespread AI adoption, including AI bias's effect 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]

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. US Department of Defense, "Artificial Intelligence, Autonomy Will Play Crucial Role in Warfare, General Says," press release, February 8, 2023, https://www.defense.gov/News/News-Stories/Article/Article/2928194/artificial-intelligence-autonomy-will-play-crucial-role-in-warfare-general-says/https%3A%2F%2Fwww.defense.gov%2FNews%2FNews-Stories%2FArticle%2FArticle%2F2928194%2Fartificial-intelligence-autonomy-will-play-crucial-role-in-warfare-general-says%2F.
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. "Using Artificial Intelligence to Promote Equity in Home Mortgage Access," *Edge PD&R*, 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.

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. In each section we offer two levels of depth to support both readers who want only a basic understanding and those reading for greater depth.

Note that AI is enabled not by one technology, but by a diverse "constellation of technologies."[12] 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:

- The intellectual forefathers of AI framed AI as the goal of manufacturing systems that resemble the human mind. In this normative sense, AI is a goal or aspiration that guides system design. In a descriptive sense, AI is commonly referred to as 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.
- AI uses technologies such as machine learning.
- Most AI systems are best conceived of as advanced inference engines. These inferences are used to produce predictions, 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.
- All AI systems currently in use are focused on specific applications. The pursuit of a more generalized AI is the goal of a sliver of ongoing AI research.
- 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 trained with data stored in a databank or collected in real time.

LEVEL 1 UNDERSTANDING

"A fundamental problem in artificial intelligence is that nobody really knows what intelligence is."[13]

—Shane Legg and Marcus Hutter

This section discusses the basics of AI, its benefits, system flexibility, and the way it works.

Basics of AI

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:

> "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."[14]

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

Intelligence. First, note that this definition does not explain the goal of this technology. The reason: the goal is in the name. As observed earlier, AI is normatively a *goal* enabled by a set of technologies. The bounds and aims of this goal are naturally murky because there is little consensus on what constitutes "intelligence." Although a small slice of the field is seeking to produce human-level intelligent systems, most engineers are simply trying to automate complex tasks. Some in the field believe serious research should ignore or downplay efforts to mimic human intelligence or describe AI systems. Mimicking human intelligence nevertheless has been the goal of AI's founding fathers, and most watershed moments in AI history such as AlphaGo's mastery of Go involve outmatching 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 distract from the real progress and pitfalls of most systems being engineered today that do not aim to match humans. These systems are designed for a variety of tasks and applications, both big and small. Tweet Hunter's tweet generator, for instance, has a narrow use case. It is not trying to create human intelligence; it is trying to generate human-quality tweets.[15] Regardless of their aims or applications, modern AI systems are united by a general effort to "automate intellectual tasks normally performed by humans,"[16] 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 referred to as **inference**. Inference is at the core of most if not all AI systems, and the goal of AI systems can be generalized as the goal of making good inferences. When one asks Alexa to play a song, it infers a song title based on the

sound of your words converted into code such that it can compare that title against coded titles in its database, and then it picks the most likely match.[17] Similarly,

- identifying a picture means inferring the correct match between the input picture and a given label, and
- 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.

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."[18] This is the **AI triad**:[19] algorithms, data, and microchips. These are the core input technologies that *together* enable AI. An essential theme of this introduction to AI is that each of these technologies is equally necessary because they are 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.[20] 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. 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, identifying the individual preferences of millions of music listeners or TV viewers.

System Flexibility

Today, nearly all AI systems could be categorized as **artificial narrow intelligence**,[21] designed to perform a specific, limited function.[22] 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 complex, albeit still limited, systems, such as driverless cars. Even within these narrow domains, AI can still suffer inflexibility. The more a system can deal with the unexpected corner cases in its domain, the higher its quality. Imagine a driverless car that is highly accurate, but only when road conditions are good. A driverless car could perform perfectly in most conditions, yet when it meets the 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 all narrow in scope, efforts are underway to develop so-called **artificial general intelligence** (AGI), with "the ability to achieve a variety of goals, and carry out a variety of tasks, in a variety of different contexts and environments."[23] This category represents the science fiction vision that many readers hold of AI. Note that *generality* does not imply *quality*. Just as a lion and a human vary wildly in intelligence, it is possible for AGI systems to perform general tasks at varying levels of proficiency.[24] Also note that AGI does not imply human-like AI; AGI can be as advanced as humans without necessarily mimicking our cognition.[25] 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**, an AI system that is smarter than humans in almost every domain.[26] These variations on advanced AI systems do not yet exist, and they represent only a fraction of AI R&D. To reiterate, most AI in use and development today does not hold these aims. Still, AGI investment is growing; a 2020 survey identified 72 active AGI R&D projects spread across 37 countries.[27] 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."[28] Although this term sounds 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 as one might follow a recipe. The term is often used interchangeably with **computer program** and **software**.

  Although this study 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 machine learning algorithm, the model is the software configuration that, once fed new data, can produce inferences, make predictions, and make decisions. The model is the inference algorithm, which is iteratively refined through machine learning, and thus continuously updates its configuration after processing new data.[29] 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 it stored or real-time data. This learning process is called **training** and is a necessary step to build artificially intelligent systems. In chapter 6, "Algorithms," the way this process works is explained in greater detail.

LEVEL 2 UNDERSTANDING

This section discusses the assessment of AI quality, assessment accuracy, and benchmarks.

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 measurement criteria used are highly specific to each application. This study offers a starting point, describing several common metrics and the way to approach these figures with a critical eye.

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 upon 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.[30] 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 the acceptable level of accuracy must be higher given the danger.[31] 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.[32] Other AI benefits must also be weighed against accuracy. Perhaps driverless cars could more 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 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 the fact that any metric used to evaluate a system will carry tradeoffs. As an illustration, there is often a tradeoff between measuring false positives and false negatives.[33] Choosing which to prioritize in evaluation depends on context and systems goals.

Returning to the disease-detecting AI example, suppose one is doing United States Agency for International Development aid work, and 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 one can 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.[34] 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.[35] 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**, which assesses how well the model minimizes *both* false negatives and false positives.[36]

These example metrics are widely used to assess AI that seeks to classify data, however, that is only one slice of evaluation and not necessarily ideal for all applications. Consider, for instance, how one might assess the quality of art generation software. Such a task is naturally fuzzy and, in many cases, might depend 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—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.[37] 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 that can allow researchers to compare and track results of models and determine state-of-the-art performance on a specific goal or task.[38] These benchmarks are often tailored to specific tasks. For instance, ImageNet is a popular benchmark for assessing image detection and classification.[39]

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.[40] The implication is that 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.[41]

3. AI POLICY CHALLENGES

Before digging into the technology that makes AI possible, we must first establish what AI policy looks like today and what issues are at stake. Currently, there is little in the way of artificial intelligence (AI) law and policy. Only a handful of federal laws relate 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.[42] Further, existing laws treat AI in a general sense rather than any application's specific issues. Executive action on AI is also in introductory stages. A 2019 executive order, Maintaining American Leadership in Artificial Intelligence, acts as the guiding document of American AI strategy, focusing on high-level policy, including international cooperation, technical standards, economic growth, R&D, and talent.[43] Building on this is a 2020 executive order, Promoting the Use of Trustworthy Artificial Intelligence in the Federal Government,[44] and a 2022 policy announcement from the White House, Blueprint for an AI Bill of Rights.[45] Both of these documents work to define broad guiding principles for the use of AI and AI policy. Beyond these initial policy salvos, however, there are few federal regulations of deeper substance or application-specific nuance.

At the state and local levels, policy is varied and often more application specific. In many states, there has been some limited movement, though action has largely been targeted at limited-scope and well-publicized AI applications. These applications include deepfakes,[46] autonomous vehicles,[47] and AI-assisted hiring legislation.[48] At the state and local levels, there is clearly a desire to implement regulation and manage some of the negative effects of AI, but action has targeted only issues that have been around for years. The issues that AI creates will not always be well publicized and as limited in scope.

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.

Research, Development, and Manufacturing

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.[49] This bill follows a long history of public engagement with this sector. While this 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 certain innovations and alternative designs?
3. How can policy play a role in ensuring that American industry competes with China's considerable state-led AI investments?

Algorithm development. Algorithm development and deployment has long been intertwined with public research support and policy. Early neural networks, for instance, were first introduced by the Office of Naval Research.[50] The Defense Advanced Research Projects Agency's (DARPA) Grand Challenge, a military-sponsored desert race, sought to incentivize autonomous vehicle progress through a competition and cash prize.[51] Some argue that this race supercharged autonomous vehicle breakthroughs. Although this public history of AI algorithm development is perhaps impressive, one should note that all industrial policy involves tradeoffs and risks. Policymakers should consider the following:

1. How will public investments crowd out private funding or distort research outcomes?
2. How can one best incentivize development while minimizing market distortions?
3. How can one ensure continued American AI leadership in algorithm development writ large?
4. How can algorithms be developed and designed to support democracy, freedom, and fairness?
5. What types of AI and applications should the public support? Should industrial policy focus on foundational or applied research? For military research, how does one ensure that innovations are designed for dual use?

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 demon-

strate 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 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. Of course, an entirely different question is whether this sort of model, regardless of fitness, should ever be used by courts.

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 open trade with new markets to ensure access to these goods?
2. How can the United States liberalize global trade to 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 counter China's market takeover of required rare earth metal deposits?

Talent and immigration. AI development requires a range of highly technical and highly 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 specialization in AI-related fields? What are their tradeoffs?
2. Can corporations be incentivized to provide training and apprenticeship programs to reduce educational burdens?
3. How can immigration policy be reformed to attract and retain global talent?
4. How can AI education balance technical skills with a need for free and creative thinking?
5. How can nontechnical fields be upskilled with AI knowledge to prepare those fields for the potential effect of AI?

Data resources and privacy. Policymakers should understand the scale of data used in AI, because many AI policy concerns revolve around **big data**. The scale of these data belies several important policy questions and challenges, including 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?

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 record-keeping differences across systems.[52] 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?

Quality Control

Explainability. Because AI systems focus on prediction 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 does the government know if autonomous vehicles are safe?
5. How does the government know that statistical AI models are producing high-quality predictions and results?

Bias and auditing. The data used and the bias embedded in AI algorithms can lead to incorrect or harmful results. AI-powered pulse oximeters have been found significantly more inaccurate for dark-skinned patients.[53] This issue can cause harm. In another case, Amazon found unintentional bias embedded in its hiring algorithm, which favored male applicants far more than female ones.[54] One 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,[55] as does the compute-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 byproducts, 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. Is there a Coasian approach to managing AI externalities? Or is it just a matter of minimizing the regulatory burden of controlling emissions and other externalities of data centers?
4. What waste and recycling standards and policies can ensure that waste is properly managed?

Labor disruptions. The advances in automation that flow from AI may disrupt the workforce and displace certain professions. For instance, in the United States, there are more than 3 million truckers, a generally low-education profession that could be eliminated by driverless vehicles.[56] Other industries may feel similar strains. Although there is no guarantee that AI will lead to fewer jobs, some will likely have to find new employment. As such, policymakers should consider the following:

1. How can education policy be used to upskill or reskill displaced 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. How can the government or private actors ensure redundant human skills in fields automated by AI?

Security and Safety

Cybersecurity. The interdiction of AI naturally comes with a transformation of the cyber-threat landscape. New threats can be found in AI. The massive depth and width of modern neural nets 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 vulner-

abilities 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. Further, it can be used to hunt and exploit vulnerabilities without human involvement. Conversely, AI can be used to detect intrusions 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?

Supply chain security. The supply chain that supports semiconductors is long, complex, and brittle. Chips are often manufactured abroad, leaving them vulnerable to foreign influence. This creates novel threats to American systems. 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?

Lethal autonomous weapons systems. AI algorithms make robotic weaponry that can select and engage targets without humans in the loop a reality. This is no longer science fiction; such systems are already in use on the battlefield.[57] 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 autonomous weapons justify their actions or explain life-or-death decisions?

INCOMPLETE AND EVER EVOLVING LIST

This list is not comprehensive but a small selection of the issues at stake. The hope is that this starting point can help readers understand the importance of this technology and its relationship to a broad array of policy domains. As they proceed to 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.[58]

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 systems and their governance.
- The data used to train and operate systems are the result of human curation, labeling, and cleaning.
- Human curation of data can lead systems to reflect biases. Some systems may perpetu-

ate negative biases, whereas some might be more objective.

LEVEL 1 UNDERSTANDING

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 **3-Vs**.[59, 60]

Data Volume

"We don't have better algorithms. We have more data."
—Peter Norvig, Director of Research at Google

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, conceived on the premise that progress in AI image recognition was a matter of more data, not improved algorithmic design.[61] 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, in just five years.[62] These 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, the volume of data is likely orders of magnitude larger. Estimating how much data is enough is a moving target and heavily depends on the application complexity,[63] model size,[64] accuracy requirements, and other goals. Progress has been made toward defining the relationship between algorithms and data requirements;[65] however, current models are still speculative.[66] In practice, engineers often depend on soft rules of thumb rather than empirically tested processes.[67] Today's AI engineering is more an art than a science.

Trending against big-data approaches are the increasingly common **small-data** strategies.[68] 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; **artificial data**, where representative yet fake data are synthetically created;[69] and **Bayesian methods**, where models are coded with prior information that provides problem context before learning begins, thereby shrinking the overall learning challenge.[70]

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.[71] The system bested all existing big-data-trained systems, challenging the assumption that more data is always better. Although AlphaZero's design is not universally applicable, it demonstrates the potential of small-data AI to transform future AI development.

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 are 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.[72]

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.[73] 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.

Velocity

Velocity refers to the speed "in which data is generated, distributed, and collected."[74] In general, this speaks to an AI system's ability to manage the data it needs for optimal performance.

Data generation and collection depends 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. Companies such as Facebook 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.[75]

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."[76] 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. IoT is a relatively new phenomenon, and these devices may grow in importance to AI as they are able to collect a wide variety of previously inaccessible data.[77]

Success can be contingent on distribution, a function of a web of storage and networking technologies, which are important components of many AI systems. These devices include not only **data warehouses**—large, centralized warehouses holding hundreds of servers on which vast lakes of data are stored[78]—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.[79]

Data Management

Data management dominates AI design. In fact, engineers frequently cite that data preprocessing accounts for 80 percent of engineering time.[80] The reason for this is that data are often disjointed, messy, and incomplete. Before a model can be trained, **data cleaning**, often by hand, is required to ensure that it will be usable.[81] To prepare data, engineers must often decide whether to remove outliers, weed out irrelevant information, add labels, and ensure that the data are well organized. Various methods and rules of thumb have also been developed to help fill in data gaps as needed.[82] 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 an AI knows that term. All these labels must be affixed by hand.[83]

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, language-processing AI trained on news articles may take on and perpetuate the societal stereotypes embedded in the language and viewpoint of those articles.[84] 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 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 or less important.[85] As 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 merely because it was easy or cheap to access, not because of its superior quality. The resulting AI system will then take on the qualities of that set, whatever they may be.[86] 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.[87] 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.[88]

LEVEL 2 UNDERSTANDING

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.

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.[89] 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 alerted image in a training dataset caused a classification system to misclassify thousands of images.[90] 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**, data inputs designed to trick AI systems during inference, can cause models to produce incorrect predictions.[91] 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.[92] Similar attacks have been developed for a range of other AI applications.

Note that beyond these prominent examples there are many attacks, and this new field of study is constantly changing. Mitigating and preventing these vulnerabilities will prove a major challenge as AI capabilities improve and grow widespread.

Data Standards and Data Capture

Much of the data that are collected and used are constrained or guided by **data standards** set by industry or government.[93] 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.[94] 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.[95] As a result of these policy choices, the inventory data that are recorded can vary substantially.[96] 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 and dramatic in others.

Standards dictate not only the content of data, but also the structure of their digital representation. MP3, PDF, and others will be familiar. 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 the 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.[97] 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.[98] 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.[99] 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 their speed that enables quick computation and, by extension, intelligence. This chapter 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 fast, doubling their processing speed roughly every two years. This geometrical pace is, however, not sustainable in time, and chips are reaching their physical limits. Future AI models will depend on existing semiconductor-based chips unless an alternative technology emerges to provide faster hardware.
- Microchip design and manufacturing is complex and is supported by a wide range of disciplines, technologies, and companies.

LEVEL 1 UNDERSTANDING

This section discusses microchip basics and AI chips.

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.[100] 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 both shrink in size and grow in density, that is, more transistors fitted in 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.[101] Moore made a speculative prediction that nevertheless became an organizing principle for the semiconductor industry; as a consequence, his law became a self-fulfilling prophecy. 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."[102] For data, this has built the storage capacity needed to support big data.[103] Transistors, however, are shrinking to their physical limits, and their performance no longer will advance as fast, if at all. Future improvements in chip function, and by extension AI, will require innovation beyond shrinking the transistors inside microchips.[104]

AI Chips

The past stability in the rate of growth of processing power meant that AI research focused on algorithms, sidelining discussion of hardware. To meet processing demands, researchers are turning to **AI chips** (also called **AI accelerators**), a range of chips that are designed not for a general purpose, but specifically for the unique processing needs of AI.[105]

The core advantage of AI chips is rooted in speed. **Central processing units** (CPUs), the general-purpose chip used for AI before the emergence of AI chips, are flexible but less efficient when processing AI-specific calculations than AI-dedicated chips.[106] CPUs preform inefficiently when operations are repeated in bulk and when memory is frequently accessed, requirements of most AI algorithms.[107] AI chips can solve these problems.

In brief, in addition to CPUs, there are currently three types 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 design flexibility at the GPU end and greater speed at the ASIC end, with FPGAs standing in the middle (figure 5.1).[108]

GPUs are limited-purpose chips originally designed for graphics processing but have been appropriated for AI.[109] Running a neural network, perhaps the most common AI model, requires large-scale and frequent matrix multiplication, a simple yet time-consuming mathematical operation.[110] GPUs are designed with many matrix

FIGURE 5.1: THE SEMICONDUCTOR SPEED-FLEXIBILITY TRADEOFF

multiplication units that can execute multiple operations simultaneously, a quality known as **parallelism**.[111]

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 both are custom built *and* can be updated as needed. Meanwhile, ASICs are custom built but cannot be updated.[112] FPGAs, owing to their programmability, carry certain efficiency costs. ASICs are perfectly tailored to an application's specific needs, giving them greater speed.[113] Although GPUs command a large share of the training market given their more flexible functionality,[114] a growing trend in AI inference chips is a steady gain of market share by ASICs.[115]

LEVEL 2 UNDERSTANDING

This section discusses microchips in detail and chip design and manufacturing.

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.[116] 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, this 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.[117] 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, SRAM (static random-access memory), a type of computer memory, uses a small collection of linked transistors to trap energy, thereby storing the data that energy represents.[118]

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.[119] 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 related 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 AI algorithms and data management. The scientists who design these AI chips tend to be electrical engineers by trade; algorithms and data are the specialty of software engineers.[120] Further, manufacturing requires an even more distinct skill set 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.[121] AI innovation is not the domain of just computer science.

Development and Fabrication

Development and fabrication. Microchip development goes through several core phases. To design a chip, engineers wield **electronic design automation** software that allows them to map a chip's execution units and arrange transistors.[122]

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.[123] The long list of materials required spans a large portion of the periodic table. Therefore, manufacturing requires an extensive supply chain, materials stock, and chemistry knowledge base to support manufacturing operations.[124] 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 UV (ultraviolet) rays, rather than visible light, in photo lithography can embed chips with thinner components, decreasing chip size and increasing chip speed.[125] To reiterate, AI innovation is not the domain of only computer science.

As a generality, the equipment used in development and manufacturing is highly specialized and, as a result, highly expensive. Photolithography scanners, for instance, can cost more than $100 million per unit.[126] Specialization has also led to concentration. In some cases, this concentration is geographical; for example, 85 percent of leading-edge chips are manufactured in Taiwan and the remaining 15 percent in South

Korea.[127] 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.[128] All of these factors complicate the robustness and security of the AI supply chain and have recently received significant policy scrutiny.[129]

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.[130] 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.[131] 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 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. Through learning, models improve their performance through iterative data analysis; this function is known as "training the model" 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.

Algorithms have the following characteristics:

- 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 of 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 underway 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.

LEVEL 1 UNDERSTANDING

This section addresses the basics of AI algorithms, machine learning, and associated technology and policy.

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, thus giving the impression that the machine is *learning*. Once inferences reach performance goals, the machine can be put to practical use, inferring on new data. Notably, models are not fixed; learning often continues after an AI model is put to practical use.

This section focuses on this dominant algorithmic technique for developing AI models—**machine learning**. Although other tools are used to create AI models, 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 artificial intelligence.

To create an AI system, engineers must select a machine learning algorithm. The type of machine learning algorithm used 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.[132] 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 needing human terms and understanding. For example, supervised learning can teach algorithms to label images of fruit with their correct English name. 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. Unlike supervised learning, which learns to perform discrete and human-defined tasks, unsupervised learning takes in unlabeled data, sifts through them, learns what hidden patterns and features they contain, and then clusters this information according to found categories.[133] This approach is useful in data analysis where humans are prone to missing important data features and overlooking unobvious correlations. 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 human-provided lenses, however, can be a challenge. Although an unsupervised algorithm can categorize data, it might not understand how to define its discoveries in human terms or match them to human objectives.
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.[134] 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 by their shapes. Then to label these groupings, the AI can learn their names using a human-provided answer key.[135] The result is an AI model that can recognize the alphabet.

4. **Reinforcement Learning.** Reinforcement learning is driven by process rather than 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.[136] 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.[137] 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. The challenge is that sometimes AI can cheat by following strategies misaligned with human goals. For example, if the goal were to maximize fuel efficiency of 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* of data, and reinforcement learning produces models that yield *actions* to take on the basis of data.[138]

Learning and Inference

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

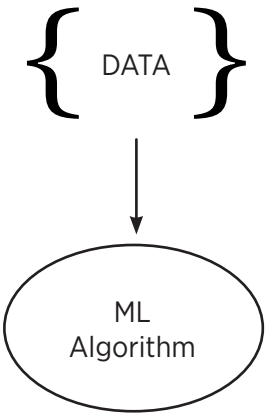
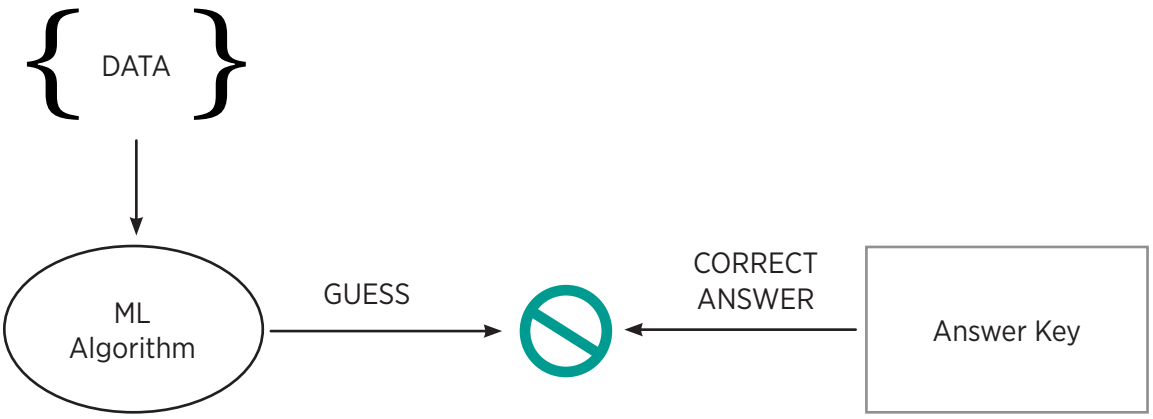


FIGURE 6.2: BENCHMARK MODEL



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 uses 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 one's 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.

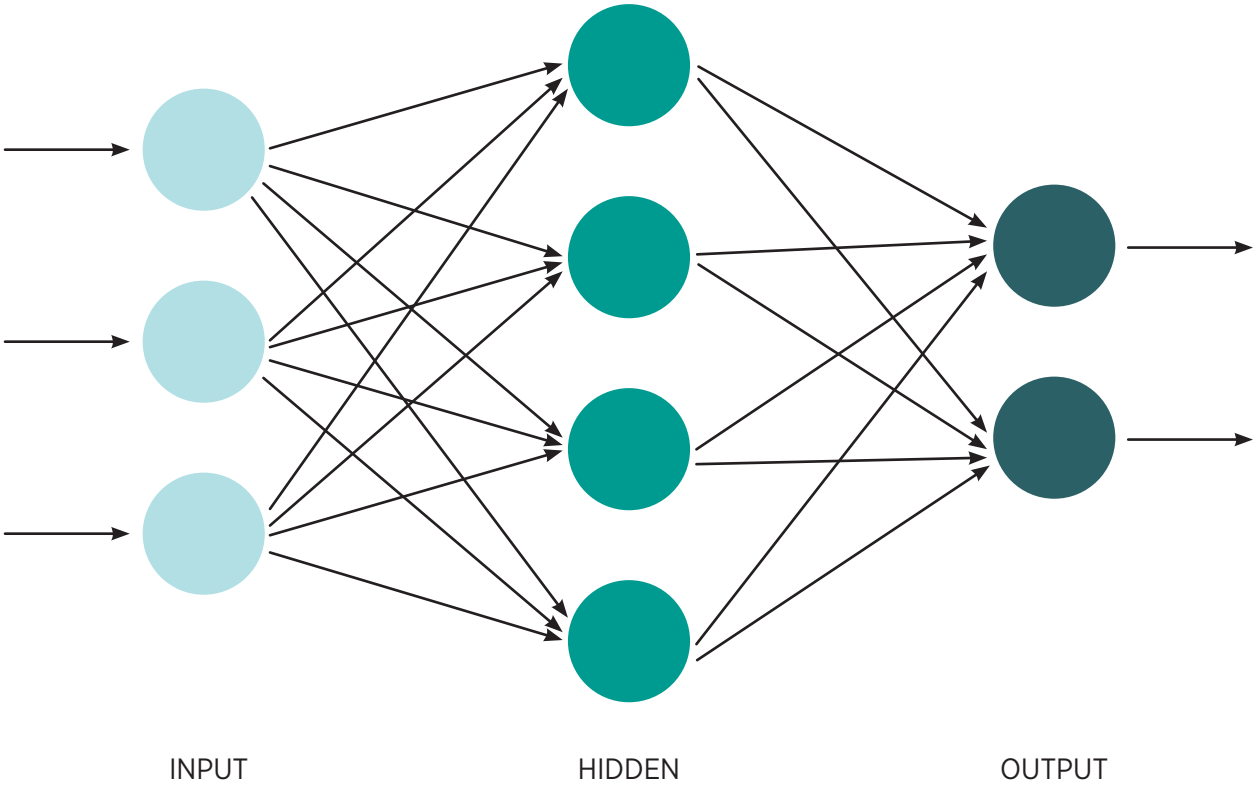
When the label and prediction differ, this incongruity signals to the model that it must change. Guided by a mathematic process, the model then gently tweaks certain internal settings and knobs called **parameters**, which are the values that shape its analytical processes. These tweaks ought to improve the model's predictive abilities for future trials. Note that although guided by mathematics, these tweaks do not guarantee improvement.

Finally, 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 the model 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** (figure 6.3). This work uses neural networks to illustrate AI inference because they are behind most modern AI innovations, including driverless cars, AI art, and AI-powered drug discovery. Just as machine learning has become synonymous with AI, many often treat neural networks as synonymous with machine learn-

FIGURE 6.3: ARTIFICIAL NEURAL NETWORK



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

ing. 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;[139] and clustering algorithms, which seek to sort data into clusters based on various metrics of data similarity.[140]

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 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 if an image were a face, a freckle might receive a low weight because this feature is not highly indicative of a face; it 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.[141] These weights are one 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 this 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**,[142] a term that describes the multiple layers of neurons that data must pass through before a final prediction can be made.[143]

KEY CHALLENGES

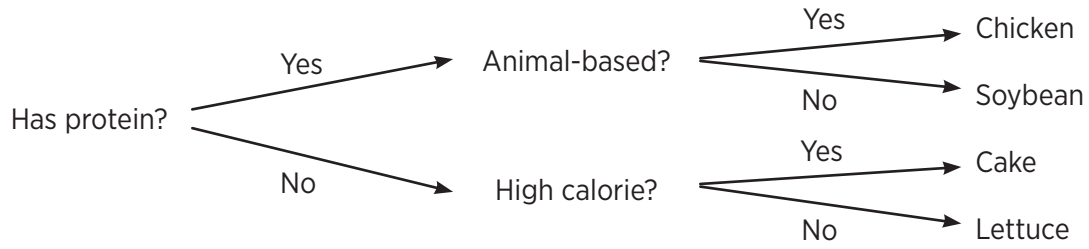
The key challenges of algorithms are model bias, explainability, and auditing of AI.

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

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.

what goal this decision will further.[144] 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 bucket individuals. If the firm seeks to maximize profit, 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.[145] Other such variations in 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.[146]

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 referred to as **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 AI**, a field that involves either designing inherently interpretable machine learning models whose decisions can be explained[147] or building tools that can explain AI systems.[148]

Some classes of **inherently interpretable** models exist today. For instance, decision trees, models that autonomously create "if–then" trees to categorize data, can be visually mapped for users (figure 6.4).[149]

Inherently interpretable models, however, are limited in accuracy and scale. For models that are not inherently interpretable, as are most neural networks, analytical tools exist. An example are tools that can determine what features in the input data were most significant in determining the model's conclusions.[150] The field is underdeveloped, however, and cannot provide model-wide explanations, explain correlations between features, or produce necessarily understandable explanations.[151]

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. Explainability can also help maximize policy effect. Knowing how a geological analysis AI is producing decisions could allow officials to modify its code, pruning variables falsely correlated with the outcome.

Auditing of AI

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.[152] This finding demonstrates that consensus has not been reached on the exact benchmarks and principles that might be used to audit AI.

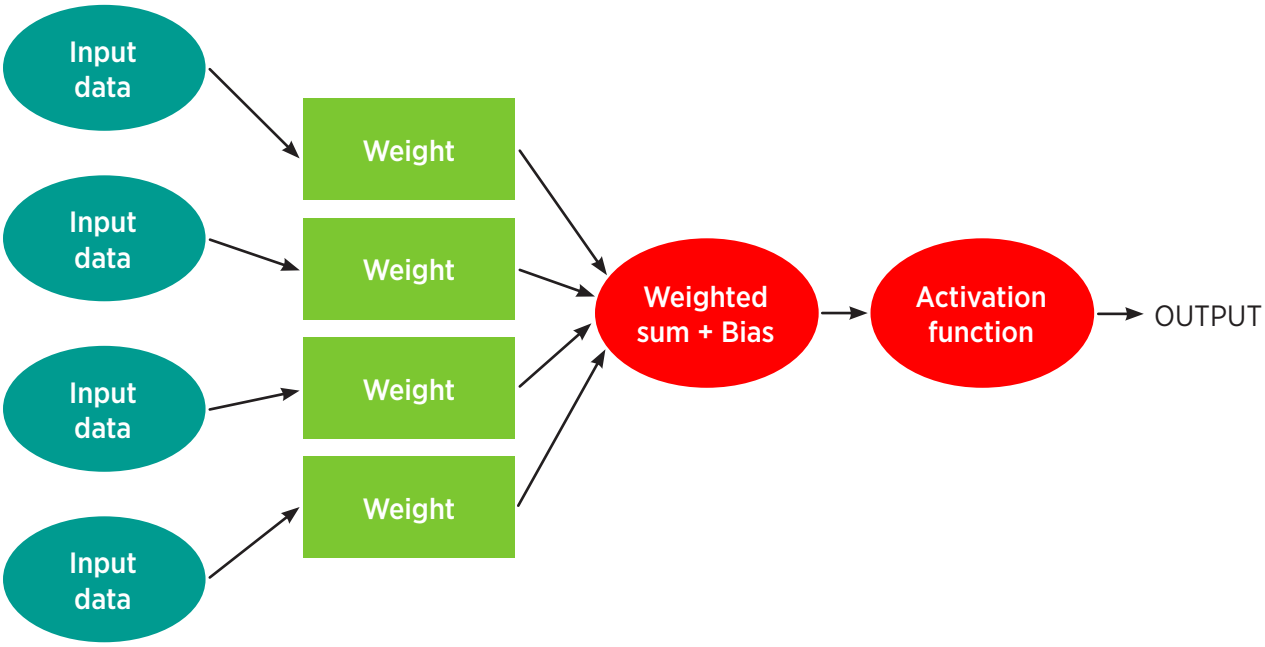
Process is another challenge. As a relatively new concept, AI audits lack frameworks and best practices, and commentators have noted that research on testing, evaluation, verification, and validation of AI algorithms has not kept pace with other subdomains of AI innovation.[153] Current processes and 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 effects or to simply eliminate data extraneous to engineering goals. Black box testing, where 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.[154] 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, during testing, and continuously after deployment.[155] 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 naturally differ from those of a music-generation AI.[156] Deter-

FIGURE 6.5: MODEL OF A PERCEPTRON, A FORM OF ARTIFICIAL NEURON



mining the questions that must be asked, the processes to be 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.

LEVEL 2 UNDERSTANDING

The previous section discusses 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 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 together. 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.[157] The activation function compares the manipulated data to certain criteria, which dictate the final output that the neuron will produce. In our 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 this result is the neuron's final decision, note that in complex neural networks this 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; these represent 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 is 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.[158] 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

This section puts together 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 together 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 of 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**[159] 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.[160]

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 where 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 where 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 exhibit diminishing returns on predictive performance.[161] 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 an engineer does not have enough data, he or she 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.[162] 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 that 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. The engineer must choose the best model for his or her 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 programing language.[163]

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. These systems are not perfect, nor are they scientific. They can, however, still produce highly accurate results.

Model Variety

The neuron illustration presented specifically a **feed-forward neural network**, a classic form that takes in data and directly maps them to a specific output.[164] For the prime-subprime categorization 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 both 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.

Generative adversarial neural networks (GANs). A GAN is a training model that uses two separate neural nets that compete against each other to learn and improve. One produces fake data trying to trick the other model into misclassifying them as real, while the other is competing to improve its abilities at distinguishing these 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.[165] 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 or 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.[166] 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."[167] As data flow through an RNN, not only are they analyzed on their own merits, but also their qualities are knit together and com-

pared 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. A model class that arrived in 2017, it has since been widely applied to 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.[168] 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. The basis for many **foundation models** today is transformers.

7. CONCLUSION: THE POLICYMAKER'S CHALLENGE

While the goal of this introduction to 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 minimizing bias, and a third perhaps would prioritize liberalism. Such general goals only 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 will require 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 both 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 data: Data that are artificially created but still thought to be generally representative of a problem. Training AI on artificial data can supplement real-world when data are poor.

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."[169]

Artificial narrow intelligence: AI built for a narrow purpose such as a specific application. This AI can do one or a few tasks 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 that data that they have learned are significant during the training process.

Bayesian methods: Models that are coded with prior information that provides context and shrinks the overall learning task and, by extension, the needed 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 hopefully capture and learn all the needed 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 artificial intelligence are represented in this form.

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

Black box: A term that refers to the often-opaque decision-making processes behind deep neural networks.

Byte: A data unit the size of 8 bits.

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

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 networks (CNNs): 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: 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 AI 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.

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 application-specific integrated circuits, because their ability to be rewritten comes with certain speed costs.

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

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.

General-purpose technology: Innovations that "[have] the potential to affect the entire economic system."[170,171]

Generative adversarial neural 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 means specifically they use pattern recognition to predict how to best generate novel output content such as images.

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.

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 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.

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.

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.[172]

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.

Overfitting: A situation where 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: 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 is placed in 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.

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, integrated circuits, 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 that data, and compares that prediction to a labeled answer key. If the inference is incorrect, the algorithm adjusts itself to improve performance.

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 Vs: 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. And *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 enable engineers' discrete control over electric currents.

Underfitting: A situation where 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.

NOTES

1. Rishi Bommasani et al., "On the Opportunities and Risks of Foundation Models," arXiv, revised July 12, 2022, https://doi.org/10.48550/arXiv.2108.07258.
2. Bommasani et al., "On the Opportunities."
3. 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.
4. Elhanan Helpman, ed., *General Purpose Technologies and Economic Growth* (MIT Press, 2003), https://mitpress.mit.edu/9780262514682/general-purpose-technologies-and-economic-growth.
5. Emily M. Bender et al., "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?," in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (New York: Association for Computing Machinery, 2021), 610–23, https://doi.org/10.1145/3442188.3445922.
6. Jacob Helberg, *The War of Wires: Technology and the Global Struggle for Power* (New York: Avid Reader Press/Simon & Schuster, 2022), 60–61.
7. "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.
8. Belinda Teoh, "Art Made by AI Wins Fine Arts Competition," *Impakter*, September 13, 2022, https://impakter.com/art-made-by-ai-wins-fine-arts-competition.
9. 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.
10. 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.
11. National Security Commission on Artificial Intelligence, "2021 Final Report," March 2021, https://www.nscai.gov/2021-final-report.
12. James E. Baker, *The Centaur's Dilemma: National Security Law for the Coming AI Revolution* (Washington, DC: Brookings Institution Press, 2020).
13. Shane Legg and Marcus Hutter, "Universal Intelligence: A Definition of Machine Intelligence," arXiv, December 20, 2007, https://doi.org/10.48550/arXiv.0712.3329.
14. National Artificial Intelligence Initiative Act of 2020, Pub. L. No. H.R. 6216 (2020).
15. "AI Tweet Generator," Tweet Hunter, accessed November 4, 2022, http://tweethunter.io/generate-tweets.
16. François Chollet, *Deep Learning with Python*, 2nd ed. (Shelter Island, NY: Manning Publications, 2021).
17. 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.

18. Ben Buchanan, "The AI Triad and What It Means for National Security Strategy," *Center for Security and Emerging Technology* (blog), August 2020.
19. Buchanan, "The AI Triad."
20. This new class of automation has raised many questions about 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.
21. Narrow AI is often alternatively referred to as brittle.
22. Ariel Conn, "Benefits and Risks of Artificial Intelligence," Future of Life Institute, November 14, 2015, https://futureoflife.org/ai/benefits-risks-of-artificial-intelligence.
23. 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.
24. Fitzgerald et al., "2020 Survey of Artificial General Intelligence Projects."
25. Fitzgerald et al., "2020 Survey of Artificial General Intelligence Projects."
26. Nick Bostrom, Superintelligence: Paths, Dangers, Strategies, 1st ed. (Oxford, UK: Oxford University Press, 2014).
27. Fitzgerald et al., "2020 Survey of Artificial General Intelligence Projects."
28. "Algorithm," *Merriam-Webster.com Dictionary*, accessed November 4, 2022, https://www.merriam-webster.com/dictionary/algorithm.
29. Jason Brownlee, "Difference Between Algorithm and Model in Machine Learning," *Machine Learning Mastery* (blog), April 28, 2020.
30. 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.
31. 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.
32. Melissa Bauman, "Why Waiting for Perfect Autonomous Vehicles May Cost Lives," *The RAND Blog*, November 7, 2017.
33. "Practices of Science: False Positives and False Negatives," University of Hawaii, accessed November 3, 2022, https://manoa.hawaii.edu/exploringourfluidearth/chemical/matter/properties-matter/practices-science-false-positives-and-false-negatives.
34. 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.
35. Bressler and Tanor, "A Guide to Evaluation Metrics."
36. 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 (AUC), receiver operating characteristic curve (ROC), mean squared error (MSE), mean absolute error (MAE), and confusion matrices, among other useful metrics. Policymakers should consider their pur-

poses, the needs of certain applications, and which metrics are best suited for those needs.
37. Ali Borji, "Pros and Cons of GAN Evaluation Measures," *Computer Vision and Image Understanding* 179 (2019): 41–65.
38. Inioluwa Deborah Raji et al., "AI and the Everything in the Whole Wide World Benchmark," OpenReview.net, modified January 14, 2022, https://openreview.net/forum?id=j6NxpQbREA1.
39. "ImageNet Large Scale Visual Recognition Challenge," ImageNet, accessed November 8, 2022, https://www.image-net.org/challenges/LSVRC/index.php.
40. Ben Dickson, "Why We Must Rethink AI Benchmarks," *TechTalks* (blog), December 6, 2021.
41. Raji et al., "AI and the Everything in the Whole Wide World Benchmark."
42. National Artificial Intelligence Initiative, "About Artificial Intelligence," accessed November 1, 2022, https://www.ai.gov/about.
43. Federal Register, "Executive Order 13859 of February 11, 2019: Maintaining American Leadership in Artificial Intelligence," National Archives, February 11, 2019, https://www.federalregister.gov/documents/2019/02/14/2019-02544/maintaining-american-leadership-in-artificial-intelligence.
44. Federal Register, "Executive Order 13960 of December 3, 2020: Promoting the Use of Trustworthy Artificial Intelligence in the Federal Government," National Archives, December 8, 2020, https://www.federalregister.gov/documents/2020/12/08/2020-27065/promoting-the-use-of-trustworthy-artificial-intelligence-in-the-federal-government.
45. The White House, "Blueprint for an AI Bill of Rights—OSTP," accessed November 1, 2022, https://www.whitehouse.gov/ostp/ai-bill-of-rights.
46. 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.
47. 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.
48. "New York City AI Bias Law Charts New Territory for Employers," Bloomberg Law, August 29, 2022, https://news.bloomberglaw.com/daily-labor-report/new-york-city-ai-bias-law-charts-new-territory-for-employers.
49. Pub. L. No. 117–167 (2022), making appropriations for the legislative branch for the fiscal year ending September 30, 2022, and for other purposes.
50. 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.
51. Defense Advanced Research Projects Agency, "The Grand Challenge," accessed November 1, 2022, https://www.darpa.mil/about-us/timeline/-grand-challenge-for-autonomous-vehicles.
52. 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.
53. 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.
54. 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.
55. 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.

56. "Economics and Industry Data," American Trucking Associations, accessed November 1, 2022, https://www.trucking.org/economics-and-industry-data.
57. 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.
58. "What Is AI Inference?," ARM, accessed November 30, 2022, https://www.arm.com/glossary/ai-inference.
59. Rebecca Laborde, "The Three V's of Big Data: Volume, Velocity, and Variety," *Oracle Health Sciences Blog*, January 23, 2020.
60. The 3-Vs have steadily expanded over time, with some sources adding additional qualities such as veracity and value. Both important qualities to be sure, but this study uses the original three for the sake of simplicity.
61. 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.
62. "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.
63. "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.
64. 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.
65. Bubeck and Sellke, "A Universal Law."
66. Bubeck and Sellke, "A Universal Law."
67. 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.
68. 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.
69. The idea of using artificial data to train systems may seem odd at first. Real-world data, however, is 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.
70. 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.
71. 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.
72. 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.
73. 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.

74. "What Is Data Velocity? Data Defined," Indicative, accessed November 29, 2022, https://www.indicative.com/resource/data-velocity.
75. 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.
76. 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.
77. 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.
78. "What Is a Data Warehouse?," Oracle, accessed November 29, 2022, https://www.oracle.com/database/what-is-a-data-warehouse.
79. 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.
80. 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.
81. This tedious task is often assigned to research assistants or Amazon's Mechanical Turk on-demand crowdsourcing service.
82. Satyam Kumar, "7 Ways to Handle Missing Values in Machine Learning," *Towards Data Science* (blog), July 24, 2020.
83. 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.
84. Jake Silberg and James Manyika, "Tackling Bias in Artificial Intelligence (and in Humans)," McKinsey & Company, June 6, 2019, https://www.mckinsey.com/featured-insights/artificial-intelligence/tackling-bias-in-artificial-intelligence-and-in-humans.
85. Baker, Hobart, and Mittelsteadt, "AI for Judges."
86. 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.
87. 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.
88. 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.
89. 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 represents 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.

90. 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.
91. 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.
92. Kevin Eykholt et al., "Robust Physical-World Attacks on Deep Learning Visual Classification," arXiv, April 10, 2018, https://doi.org/10.48550/arXiv.1707.08945.
93. In many cases, and across industries, governments and government-commissioned industry regulators control standards. Standardization is deeply intwined with policy. Most standards were decided without considering the needs of artificial intelligence, 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.
94. Jason Fernando, "GAAP: Understanding It and the 10 Key Principles," Investopedia, updated June 28, 2022, https://www.investopedia.com/terms/g/gaap.asp.
95. "LIFO vs. FIFO," Corporate Finance Institute, updated January 6, 2023, https://corporatefinanceinstitute.com/resources/accounting/lifo-vs-fifo.
96. 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.
97. Yong-Yeon Jo et al., "Impact of Image Compression on Deep Learning-Based Mammogram Classification," *Scientific Reports* 11 (2021): 7924.
98. 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.
99. Not all AI uses machine learning. Deep Blue is an example of an expert system, a form of "symbolic" artificial intelligence 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.
100. *Encyclopedia Britannica*, 2022, s.v. "semiconductor," https://www.britannica.com/science/semiconductor.
101. *Encyclopedia Britannica*, 2022, s.v. "Moore's law," https://www.britannica.com/technology/Moores-law.
102. 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.

103. 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.
104. 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.
105. 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.
106. Brian Bailey, "Von Neumann Is Struggling," Semiconductor Engineering, January 18, 2021, https://semiengineering.com/von-neumann-is-struggling.
107. 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.
108. 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 the total energy usage. Emerging designs such as analog chips could purportedly slash the energy used by chips; however, these designs have yet to affect the mainstream.
109. 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. As 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.
110. 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 chapter 4.
111. Falan Yinug, "Semiconductors: A Strategic U.S. Advantage in the Global Artificial Intelligence Technology Race," 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.
112. Mike Brogioli, "The DSP Hardware/Software Continuum," in *DSP for Embedded and Real-Time Systems*, ed. Robert Oshana (Oxford, UK: Newnes, 2012), 103–12.
113. "ASIC vs. FPGA: What's the Difference?," AsicNorth, October 10, 2020, https://www.asicnorth.com/blog/asic-vs-fpga-difference.
114. 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.

115. Batra et al., "AI Hardware."
116. "Introduction to Semiconductors," AMD, accessed December 31, 2021, https://www.amd.com/en/technologies/introduction-to-semiconductors.
117. *Encyclopedia Britannica*, 2020, s.v. "dopant," https://www.britannica.com/technology/dopant.
118. "Understanding RAM and DRAM Computer Memory Types," ATP, July 1, 2022, https://www.atpinc.com/blog/computer-memory-types-dram-ram-module.
119. 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.
120. 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.
121. Hunt and Zwetsloot, "The Chipmakers."
122. Khan, "AI Chips."
123. "Semiconductor Fabrication: How Are They Manufactured?," Halocarbon, December 12, 2022. https://halocarbon.com/semiconductor-fabrication-how-are-they-manufactured.
124. Khan, "AI Chips."
125. 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/.
126. Hunt and Zwetsloot, "The Chipmakers."
127. 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.
128. Katie Tarasov, "ASML Is the Only Company Making the $200 Million Machines Needed to Print Every Advanced Microchip. Here's an Inside Look," CNBC, updated March 23, 2022, https://www.cnbc.com/2022/03/23/inside-asml-the-company-advanced-chipmakers-use-for-euv-lithography.html.
129. 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.
130. Lohn and Musser, "AI and Compute."
131. Kim Martineau, "What Is Federated Learning?," *IBM Research* (blog), August 24, 2022.
132. "What Is Supervised Learning?," IBM, 2020, https://www.ibm.com/cloud/learn/supervised-learning.
133. "What Is Unsupervised Learning?," IBM, 2020, https://www.ibm.com/cloud/learn/unsupervised-learning.
134. Jason Brownlee, "What Is Semi-Supervised Learning," Machine Learning Mastery, April 9, 2021, https://machinelearningmastery.com/what-is-semi-supervised-learning.
135. Ben Dickson, "What Is Semi-Supervised Machine Learning?," TechTalks, January 4,

2021, https://bdtechtalks.com/2021/01/04/semi-supervised-machine-learning.

136. Piyush Verma and Stelios Diamantidis, "What Is Reinforcement Learning?," Synopsys, updated April 27, 2021, https://www.synopsys.com/ai/what-is-reinforcement-learning.html.
137. M. Tim Jones, "Models for Machine Learning," IBM, December 5, 2017, https://developer.ibm.com/articles/cc-models-machine-learning/#reinforcement-learning.
138. M. Tim Jones, "Models for Machine Learning."
139. "AI Modeling: Driving Intelligence in Analytics," Intel, accessed December 14, 2022, https://www.intel.com/content/www/us/en/analytics/data-modeling.html.
140. "Clustering," scikit-learn, accessed December 14, 2022, https://scikit-learn.org/stable/modules/clustering.html.
141. 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.
142. AI, ML, and ANNs: the difference between these terms can seem fuzzy and confusing. It is often helpful to think hierarchically. At the top is artificial intelligence, which is the overriding goal of all of these technologies. Next is machine learning, 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, you can find more specify 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 large artificial neural networks, or ANNs. As machine learning and deep learning have become exceedingly popular in artificial intelligence, 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.
143. "What Is Deep Learning?," IBM, accessed February 22, 2021, https://www.ibm.com/cloud/learn/deep-learning.
144. 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-happens and-why-its-so-hard-to-fix.
145. Adam Zewe, "Can Machine-Learning Models Overcome Biased Datasets?," *MIT News*, February 21, 2022, https://news.mit.edu/2022/machine-learning-biased-data-0221.
146. 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.
147. 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.
148. 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.
149. "Decision Trees," scikit-learn, accessed December 12, 2022, https://scikit-learn.org/stable/modules/tree.html.
150. "Algorithm Descriptions," Captum, accessed December 11, 2022, https://captum.ai.

151. Narine Kokhlikyan, "Opening Up the Black Box: Model Understanding with Captum and PyTorch," PyTorch, 2020, https://www.youtube.com/watch?v=0QLrRyLndFI.
152. 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 Iniversity, May 28, 2021, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3855171.
153. 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.
154. Alex Engler, "Auditing Employment Algorithms for Discrimination," Brookings Institution, March 12, 2021, https://www.brookings.edu/research/auditing-employment-algorithms-for-discrimination.
155. Reva Schwartz et al., "Towards a Standard for Identifying and Managing Bias in Artificial Intelligence," National Institute of Standards and Technology, March 15, 2022, https://www.nist.gov/publications/towards-standard-identifying-and-managing-bias-artificial-intelligence.
156. Marietje Schaake and Jack Clark, "Stanford Launches AI Audit Challenge," Stanford University Human-Centerd Artificial Intelligence, July 11, 2022, https://hai.stanford.edu/news/stanford-launches-ai-audit-challenge.
157. In a realistic sense, the activation function isn't a trigger but the mathematical rules of the road for the type of output that 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.
158. In a more mathematical sense, the bias also serves to orient the function towards 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.
159. "What Is Supervised Learning?," IBM.
160. "What Is Supervised Learning?," IBM.
161. "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.
162. 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.
163. "Machine Learning in Python—Scikit-Learn 1.2.0 Documentation," scikit-learn, 2022, https://scikit-learn.org/stable/index.html.
164. 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.
165. Jason Brownlee, "A Gentle Introduction to Generative Adversarial Networks (GANs)," *Machine Learning Mastery* (blog), July 19, 2019, https://machinelearningmastery.com/what-are-generative-adversarial-networks-gans.
166. 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.
167. Simeon Kostadinov, "How Recurrent Neural Networks Work," *Medium*, accessed

November 10, 2019, https://towardsdatascience.com/learn-how-recurrent-neural-networks-work-84e975feaaf7.
168. Dickson, "What Is Semi-Supervised Machine Learning?"
169. National Artificial Intelligence Initiative Act of 2020, Pub. L. No. H.R. 6216 (2020).
170. Pethokoukis, "How AI Is Like That Other General Purpose Technology, Electricity."
171. Elhanan Helpman, ed., *General Purpose Technologies and Economic Growth* (MIT Press, 2003), https://mitpress.mit.edu/9780262514682/general-purpose-technologies-and-economic-growth.
172. 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.