Harvard Business School

9-324-062
REV: AUGUST 27, 2024

PAUL HEALY
DEBORA SPAR

Monsters in the Machine? Tackling the Challenge of Responsible AI

When you see something that is technically sweet, you go ahead and do it and you argue about what to do about it only after you have had your technical success.

— J. Robert Oppenheimer[1]

If we leave Google and Microsoft to do AI, the robots are going to destroy us.

— Elon Musk[2]

On November 30, 2022, OpenAI released ChatGPT 3.5, an artificial intelligence chatbot that allowed users to have human-like conversations with a machine. The application could respond to user questions, compose essays on virtually any topic, write poems in any format, respond to emails, create computer code, and perform countless other tasks. By January 2023, the software application had become the fastest-growing consumer application in history, attracting more than 100 million users and pushing OpenAI to a valuation of $29 billion.

Upon the release of ChatGPT, observers from across all sectors of society swiftly agreed that the world was on the edge of a breakthrough. Chatbots like ChatGPT had the potential to address previously insoluble problems, and to free human workers from an infinitude of mundane tasks. The release, however, sparked an intense debate about the potential risks of AI, including massive job displacements, privacy violations, copyright infringements, and all the unseen consequences a non-human intelligence could unleash. New companies like Open AI, along with established giants such as Microsoft, Google, Amazon, and Meta faced intense questions about how the new technology should be used and regulated. Should this task be left to the developers themselves or should governments take the reins? Should the public have a say, or was AI too complex to be left in amateur hands? As of 2024, AI technologies were evolving faster than anything before them ever had. Should society let products like chatbots develop on their own? Or did somebody, somewhere, need to put guardrails in place before it was too late?

Professors Paul Healy and Debora Spar prepared this case with the assistance of Amram Migdal and Julia Comeau. This case was developed from published sources. Funding for the development of this case was provided by Harvard Business School and not by the company. HBS cases are developed solely as the basis for class discussion. Cases are not intended to serve as endorsements, sources of primary data, or illustrations of effective or ineffective management.

Machine Learning and Artificial Intelligence

The idea of using computer programs to make decisions and improve predictions was first developed in the 1930s by computer science pioneer Alan Turing. The concept developed slowly over the next several decades, as computing power rose gradually to meet Turing's ambitions: although an early AI program was invented in 1955, for example, it took until 1997 for the first publicly available speech recognition software to enter the market.[3] And it wasn't until the 2010s that neural network models had sufficient computing power to begin to enable such complex tasks as extracting information from pictures (computer vision), transcribing and translating spoken words (natural language processing), discovering insights and patterns in text (natural language understanding), guiding machinery to move autonomously (robotics), and finding patterns in large data sets (deep learning and neural networks).

During the 2010s and 2020s, the development of AI consistently outpaced expert predictions and produced surprising capabilities.[4] In August 2020, for example, an AI flight simulator defeated an experienced jet fighter pilot in a virtual air battle.[5] In 2022, AI chess, Go, and other gaming programs could reliably defeat the world's best human players (See **Exhibit 2**). Encouraged by these results, a small but growing band of researchers were starting to believe that machines were on the verge of developing human-like powers of intelligence.[6] Specifically, they referred to the test—or the challenge, really—that Turing had laid out in 1950: could any machine convince a human interacting with it that it was human? Could machines "imitate" humans at their own game? By 2023, at least in some small spheres, it seemed they could.

At this stage, most applications of AI were still considerably more targeted than anything resembling the general reasoning power of human intelligence. That is, although AI could outperform humans in specific domains, no computer possessed the broad problem-solving abilities of the human brain or the many nuances of human reasoning. Many researchers, however, believed that recent breakthroughs in AI were only the beginning, and that computers could indeed develop fuller and more human-like modes of intelligence. They called these modes artificial general intelligence, or AGI.

Commercial Applications

Meanwhile, even as this research raced ahead, companies across the world were already using AI advances across their operations, changing how they did business, what they offered, and the scale of impact they could achieve. In 2016, Microsoft CEO Satya Nadella articulated the scale of AI's potential impact at Microsoft's annual Ignite conference, saying: "We are not pursuing AI to beat humans at games. We are pursuing AI so that we can empower every person and every institution that people build with tools of AI, so that they can go on to solve the most pressing problems of our society and our economy."[7]

In healthcare, AI's potential applications included improving detection, diagnosis, and screening of health risks; predicting effective treatment plans; enhancing providers' capabilities via access to expert-level knowledge; reducing error rates; and automating multiple aspects of administration.[8] Theoretically, AI systems could also supplement the limited healthcare services and capabilities available in under-served areas.[9] In finance, AI could similarly be deployed for a wide variety of purposes: detecting fraud, for example, combating money laundering, and performing complex tasks such as risk assessments, data analyses, and compliance management.[10]

Most companies that deployed AI applications, however, did not develop these applications on their own. Rather, they licensed use of the algorithms and cloud computing power from larger companies such as Microsoft, Alphabet (DeepMind, Google), Amazon, Apple, IBM, Meta, Palantir, Salesforce, and OpenAI, accessing their services via APIs (application programming interfaces). Outside the United States, Chinese players such as Alibaba, Baidu, ByteDance, Megvii, and Tencent offered similar services to their customers, as did a growing ecosystem of smaller players.[11] Globally, the AI market was worth an estimated $330 billion in 2022, with projections to increase to $1.5 trillion by 2030.[12]

Fears of the Smart Machine: Evaluating the Risks of Artificial Intelligence

Not surprisingly, concerns about abuses of AI arose almost as soon as the technology was conceived. In 1951, Turing himself observed that "at some stage [. . .] we should have to expect the machines to take control."[13] But the original predictions of rapid progress and potential doomsday scenarios proved premature. Instead, most early applications of AI were powered by relatively simple machine learning algorithms and targeted at relatively narrow problems, such as recognizing patterns in piles of data, abstracting customer insights, or optimizing inventory in supply chains.

Most experts doubted that AI would become self-aware any time in the foreseeable future, or that it would ever develop the means to threaten humanity as portrayed in films like *Terminator* or *2001: A Space Odyssey*. But the negative potential of AI was already apparent, even in commonly used technology. Companies in many industries, for example, used facial recognition AI for identity verification and access, and face blurring for privacy. Some even used early versions of the technology in ways that potentially compromised their customers' privacy, or created risks that they might be incorrectly identified—a particular threat in the area of law enforcement. Even more frightening was that, when these mistakes occurred, AI operators and creators often could not account for how the algorithms reached their conclusions—the so-called "black box" dilemma.

In 2012 and 2013, a survey conducted by Oxford philosopher Nick Bostrom found "more than half of the top 100 most cited AI researchers believed there is a substantial (at least 15 percent) chance that the effect of human-level machine intelligence on humanity will be 'on balance bad' or 'extremely bad' (existential catastrophe)."[14] As AI grew more prominent in daily life, the general public began to share experts' opinions on the threat posed by the technology. In 2023, a survey conducted by the Pew Research Center asked over 11,000 Americans whether they were more concerned or excited about the use of artificial intelligence in daily life. The majority (52%) reported that they were more concerned, while only 36% stated that they were more excited.[15] On Manifold, a site in which users bet on future events, 16% of participants wagered that AI would wipe out humanity by 2100.[16]

Conceiving the Costs

Meanwhile, even as these existential threats were being debated, several more immediate costs were becoming increasingly apparent. To begin with, because AI had the ability to replicate a wide range of both blue- and white-collar jobs, it also had the obvious potential to displace human workers who had long worked in these fields. According to a 2023 report from McKinsey, up to 30% of working hours across the U.S. economy could be automated by 2030—a trend begun in the era of COVID-19 and accelerated by the rise in generative AI.[17] To be sure, breakthrough technologies from the steam engine to the sewing machine had long displaced human laborers, and most of these laborers (in demographic terms at least) swiftly found new employment—often under better terms and with higher pay.[18] But because the technological disruption of AI promised to be both seismic and widespread, many observers feared that job losses in an age of AI might easily overwhelm any job gains, or wipe out entire segments of well-established careers.[19]

Second, because AI systems could create an infinity of images, videos, and conversations, they also could be used to spread these messages in deeply irresponsible ways—manipulating voters, for instance, fueling partisan divides, or misrepresenting scientific truth. Most foundationally, AI-generated content did not necessarily distinguish between fact or fiction, or between information created by human authors or something else. There were also no apparent property rights in these early systems, which worked largely by training their models on huge tracts of information gathered from across the Internet, and no attribution of authorship. Other risks centered around the possibility for AI systems to invade individuals' privacy or to use their personal data in unfair or deceptive ways. Both governments and corporations, for example, could theoretically use AI-enabled technologies to monitor individuals' movements and decisions without their knowledge or consent, or even track their emotional states.[20]

AI also amplified traditional data security concerns such as the risk of unintentionally leaking private data. Algorithms powering self-driving cars, for example, shared data with smart traffic signals and other public and private networks. AI-enabled medical records pulled and sometimes pooled intimate details about patients' physical, mental, family, genetic, and sexual health. Would individuals need, or even be able, to give consent each time their data was accessed by a potentially wide array of users? Would, or should, they be able to pull their personal data from AI-enabled systems, or prevent their information from being used to train increasingly sophisticated large language models?

Finally, because AI software tended to reinforce whatever biases were embedded in the datasets that trained them, many observers feared that it would perpetuate and perhaps even expand the conscious and unconscious biases of its human programmers.[21] Or as Harvard political philosopher Michael Sandel argued, "AI not only replicates human biases, it confers on these biases a kind of scientific credibility. It makes it seem that these predictions and judgments have an objective status."[22]

Early studies confirmed these fears, finding numerous cases of model discrimination and bias.[23] In Broward County, Florida, for example, only 20% of the defendants predicted by an AI model to be at high risk of committing future violent crimes were actually subsequently prosecuted for such crimes.[24] The model was also more than twice as likely to incorrectly predict future violent crimes for Black defendants than for White.[25]

Capturing the Externalities

Conceptually, many of these risks could be described as negative externalities, or what Milton Friedman and other economists occasionally called "neighborhood effects." In simplest terms, these were the unintended consequences of commercial production—things like pollution from a chemical factory that fell on the fields of a neighboring farmer, or noise from a neighborhood bar disturbing a nearby yoga studio. In these instances, the cost of production was borne by someone other than the producer or consumer, and thus could not be neatly captured in price. Or as Ronald Coase, who won a Nobel prize for his work in this area, wrote, these are "case[s] in which, although the pricing system is assumed to work smoothly (that is, costlessly), the damaging business is not liable for any of the damage which it causes. This business does not have to make a payment to those damaged by its actions."[26]

Traditionally, as Coase and others described, solutions to such problems came via one of three types of state intervention: the government could regulate the undesirable behavior (barring or limiting certain types of emissions, for example); it could tax them; or it could ban them directly. None of these were perfect, as legions of economists had subsequently argued, and all were complicated. Regulation imposed its own costs, along with the risk of stifling innovation.[27] Taxation faced obvious political obstacles, and outright bans were generally only feasible either under authoritarian regimes or in cases of obvious and unmitigated risks. As a result, even though the economics of externalities were well understood, policies for addressing them tended to fall in and out of favor, varying in accordance with the prevailing political sentiments of a state or an era.

One alternative to state intervention was self-regulation, an option championed by Elinor Ostrom, who won a Nobel prize in 2009 for her work in this area. Based largely on examples drawn from common pool resources such as fisheries or grazing land, Ostrom argued that self-interested actors (and, by analogy, firms) could be prodded to develop and enforce mechanisms for their own collective regulation. Or as she wrote, "[T]he evidence does not support the assumption that individuals always maximize expected, short-term, material returns to self."[28] Instead, under the right conditions, firms would realize the individual benefits of short-term restraint, and collectively decide not to engage in activities that could undermine their industry in the long run.

Variations of self-regulation were fairly common across the American industrial landscape. The American Bar Association had long set and enforced ethical codes for the country's lawyers, just as the American Medical Association established and oversaw codes of medical ethics.[29] In higher tech sectors, associations such as the Bluetooth Special Interest Group and the 3rd Generation Partnership Project led early efforts to establish technical standards for wireless and broadband communication, respectively; while the digital music industry was arguably brought into being through the creation of industry-led standards such as the MP3 file format.[30] Proponents of these standard-setting models argued that they were faster and more flexible than any state-led project could possibly be, and that industry insiders were better equipped to encourage the kinds of behavior that would help their sector to thrive.[31] Critics, however, pointed to the obvious flaw: that commercial firms were inherently conflicted between creating rules and standards that would benefit society, and those that would benefit their own bottom line. Or as Virginia Haufer, a leading scholar of self-regulation explained, "Industry self-regulation may be one way to raise standards, but because those standards are voluntary and unenforceable, they lack credibility. Even more troubling for many, however, is the issue of accountability. If these efforts are an indirect means for public goals to be met by private interests, then how does the public influence their content?"[32]

Open AI and ChatGPT

Founded in December 2015 by Elon Musk, Sam Altman and a handful of other tech entrepreneurs, OpenAI began as a non-profit artificial intelligence research company, dedicated to creating the world's first Artificial General Intelligence (AGI), a machine with a human-like ability to learn and solve problems.[33] Musk and Altman, who were simultaneously excited by and wary of AGI's potential, wanted to ensure the technology was developed in a way that was both safe and open-source. In its introductory blog post, OpenAI explained, "It's hard to fathom how much human-level AI could benefit society, and it's equally hard to imagine how much it could damage society if built or used incorrectly."[34] In this vein, OpenAI was created as an altruistic venture, "unconstrained by a need to generate financial return," in which "researchers will be strongly encouraged to publish their work, whether as papers, blog posts, or code, and our patents (if any) will be shared with the world."[35] The firm's primary obligation, as stated in its charter, was to humanity.[36]

By 2018, however, cracks were beginning to appear in OpenAI's model. The firm found it increasingly difficult to compete with powerful technology firms in Silicon Valley, and Musk worried that OpenAI was falling too far behind Google and its AI platform, DeepMind. According to some sources, Musk proposed that he take control of the firm, but Altman and the board rejected this idea.

Soon after, Musk parted ways with OpenAI and reneged on a promised donation, reportedly out of frustration over the leadership decision.[37] Publicly, though, OpenAI stated that Musk left to eliminate any future conflict of interest between OpenAI and Tesla, where researchers were developing their own AI for self-driving cars.[38]

After Musk's departure, OpenAI lost much of its promised funding, further widening the gap between the small non-profit and the Silicon Valley giants it was competing against. By 2019, OpenAI recognized that the non-profit model could not sustain the expensive artificial intelligence research that pursuing AGI would require. Accordingly, on March 11, OpenAI announced that it would add a new "capped-profit" entity to its organization called OpenAI LP, whose profits would be passed to the overarching non-profit organization, OpenAI Inc. "Capped-profit," the company explained, was "a hybrid of a for-profit and a nonprofit," in which investors would be able to collect returns up to a certain cap (originally set at one hundred times the initial investment). OpenAI claimed that this shift was necessary to compete with other major technology firms. About six months later, Microsoft invested one billion dollars into OpenAI.[39]

Using these funds, engineers at OpenAI built a dedicated supercomputer to train AI models, and to create what would eventually be released as the AI chatbot ChatGPT and the image generator DALL-E.[40]

DALL-E was released first: in a blog post published on January 5, 2021, OpenAI announced that it had created a network capable of generating unique images from text prompts.[41] A portmanteau of the Pixar robot WALL-E and the Spanish surrealist artist Salvador Dalí, DALL-E was trained on more than 650 million images and used captions associated with these images to generate its own artistic or photorealistic pictures. Then, on November 30, 2022, OpenAI released ChatGPT, a type of large language model capable of generating human-like speech in response to commands.[42] ChatGPT became an instant hit—within five days, the platform amassed over one million users; within two months, it had 30 million users and five million visitors per day.[43] People from all industries and backgrounds found uses for ChatGPT: everything from generating automatic responses to common inquiries in customer service fields, to analyzing and interpreting medical literature and generating video game storylines.

By May of 2023, OpenAI was valued at $29 billion and had raised $11.3 billion in funding, with an additional pledge from Microsoft to invest $10 billion over the following few years.[44]

The Valley Responds

After OpenAI released ChatGPT in 2022, many of Silicon Valley's most powerful players began racing to deploy their own AI products. Google, for example, declared a "code red" and swiftly began reassigning teams to help develop and release AI products faster, including its own search-oriented chatbot, Bard.[45] Chinese tech firms such as Baidu and WeChat owner Tencent reportedly scrambled to take advantage of ChatGPT's popularity, rolling out their own versions of generative AI platforms within months of ChatGPT's release.[46] And Microsoft announced plans to invest $10 billion in OpenAI and integrate generative AI into its Office software and search engine, Bing. "A race starts today," commented Nadella, "We're going to move, and move fast."[47] Like Altman, many of these firms expressed concern about AI's dangers even as they were building increasingly powerful AI systems of their own. And as they argued publicly for some form of regulatory action, they also began to write their own rules of the evolving game.

Indeed, over the course of 2023, each of the major tech firms created their own principles to guide the development of AI. Though the principles varied slightly, each prioritized a similar set of values: privacy and security, fairness and inclusion, transparency, safety, and accountability.[48]

Microsoft

Microsoft, the world's largest software company by revenue,[49] established its AI and Research group in September 2016.[50] Several months later, it announced a portfolio of new services under the auspices of its Azure business, designed to help developers "build intelligent apps, with understanding and natural user interaction capabilities."[51] By 2022, Azure offered a range of AI services, equipped with "high-quality vision, speech, language, machine learning models, and more,"[52] targeting industries that ranged from healthcare to finance and government services.[53] It had also become the company's most profitable segment.

As it was unrolling these services, Microsoft established its AI Ethics in Engineering and Research Committee (Aether), to ensure that any business decisions regarding AI applications complied with the firm's broader ethical principles. Specifically, Aether's mission was to review any potentially sensitive application of AI with regard to Microsoft's six stated AI values of fairness, reliability and safety, privacy and security, inclusiveness, transparency, and accountability. Aether was also responsible for considering the broader implications of AI for the company, its customers, and society in general.

Under the terms of the company's Responsible AI Standard, all new AI systems built by Microsoft had to comply with the six AI values mentioned above. Under the accountability standard, for example, any new AI system had to demonstrate that it provided a valid solution to the problem it was designed to solve and identify any potential adverse impacts. Under the transparency requirement, Microsoft AI systems had to be designed to "inform people that they were interacting with an AI system or using a system that generated or manipulated image, audio, or video content that could falsely appear to be authentic."[54]

In January 2023, however, just as interest in AI seemed to be reaching a fever pitch, Microsoft quietly changed direction, eliminating the entire Aether team as part of a broader round of layoffs. Several months later, when OpenAI's board of directors briefly tried to fire Altman, Microsoft stepped into the breach, first by offering to hire Altman and then, when he returned to OpenAI, by confirming its investment and confidence in the firm. At year's end, Microsoft had invested more than $13 billion in OpenAI and held a 49% ownership stake.

Google

Like Microsoft, Google had been pursuing AI for over a decade. It launched Google Brain in 2011, a research project led by Stanford professor Andrew Ng, that succeeded in building a deep-learning software system on top of Google's existing cloud computing infrastructure. In 2014, it purchased the British artificial intelligence research laboratory DeepMind for $400 million, folding that group's experience with game-playing systems into its expanding understanding of what was commonly called "reinforcement learning"—prodding computers to learn to improve their behavior to achieve goals.[55] By 2023, it had deployed these and similar problem-solving skills to a host of Google products, using AI, for example, to generate predictive text in Google Docs, create images for Google Slides, and power a chatbot called Bard.[56]

As these technologies were developed and launched, Google had also implemented several self-regulatory practices to guide their use. To prevent bad actors from abusing its Universal Translator, for instance, Google limited its use to authorized partners. The firm also prioritized being transparent with its use of AI, announcing in May 2023 that it would "provide ways to identify when we've used [AI] to generate content in our services."[57]

Publicly, Google's high-profile former CEO, Eric Schmidt, was appointed to the National Security Commission on Artificial Intelligence, advising on how best to use the technology to advance U.S. interests, build the workforce of the future, and develop ethical procedures. Under Schmidt's leadership, the commission wrote proposed legislation that later became law and steered billions of dollars of taxpayer funds towards advancing the AI industry. Schmidt himself made more than 50 investments in AI companies, giving himself an economic stake in the industry even as he was creating its rules for regulation.[58]

Meta

On November 15, 2022, just weeks before OpenAI launched ChatGPT, Meta (formerly known as Facebook) unveiled its own large language model called Galactica. Unlike ChatGPT, Galactica was designed specifically as a "shortcut" for scientists, a tool Meta boasted could "summarize academic papers, solve math problems, generate Wiki articles, write scientific code, annotate molecules and proteins, and more."[59] Within hours of Galactica's release, however, researchers began reporting biased and incorrect outputs from the model, and within days, the company had pulled its public demo.[60]

Nonplussed, Meta rapidly began integrating AI products into its apps in 2023. These included an AI sticker generator for Messenger, an AI-assisted image editor for Instagram, and a slew of celebrity lookalike AI chatbots (including Jane Austen, Tom Brady, Mr. Beast, and Snoop Dogg).[61] These AI chatbots, which Mark Zuckerberg described as "the future of human connection," were represented by celebrity faces, but assumed their own names and personalities.[62] Each "character" was built to discuss specific niche interests such as sports, comedy, or literature, allowing users to pick a chatbot to interact with in the same way they might pick a human friend.[63]

In creating these tools—especially those mimicking real people—Meta simultaneously implemented safeguards to prevent the spread of misinformation. The firm reported spending thousands of hours training its AI on "red-team scenarios" to try and spot weaknesses in its AI model.[64] It also implemented a number of governance best practices including evaluating products with internal and external experts, training models on safety and responsibility guidelines, and adding built-in feedback tools.[65] As AI grew more advanced, Meta believed, the technology could learn to self-regulate. Or as they stated in 2023, "We're optimistic generative AI can help us reinforce our policies in the future."[66]

Unlike Google and OpenAI, however, Meta also embraced an open-source approach to AI, making its large language model freely available to AI researchers, along with all the tools necessary to reproduce and modify the model.[67] Their argument was straightforward: open-source models, they said, were safer than proprietary ones, since they allowed more public visibility into how the models were being trained, and thus greater opportunities for public audits and opinions. "Do you want every AI system to be under the control of a couple of powerful American companies?" argued Yann LeCoun, Meta's chief AI scientist, "Progress is faster when it is open."[68] Many critics outside the firm, however, argued the open-source models opened a pandora's box of problems, potentially allowing bad actors to bypass any guardrails that might be imposed upon commercially developed models. In October 2023, a group of protesters gathered outside Meta's San Francisco office, accusing the firm of driving an "irreversible proliferation" of potentially unsafe technology.[69]

Anthropic

Anthropic, an AI safety and research company, was founded in 2021 by a group of former OpenAI employees who left the firm in 2019 following its shift to a capped-profit model and quickly raised nearly $1 billion in funding.[70]

Informed by their concern over corporate governance in the AI space, Anthropic's founders created their company as a public benefit corporation under Delaware corporate law. This structure allowed the board to legally weigh financial considerations with the public benefit purpose specified in Anthropic's founding charter. At the close of Anthropic's $450 million Series C funding, a new governance structure and class of stock—the Long-Term Benefit Trust—was created to strengthen the independence of the board and "ensure that Anthropic responsibly balances the financial interests of stockholders with the interests of those affected by Anthropic's conduct and our public benefit purpose."[71] Operationally, Anthropic's engineers sought to train their models on a set of clear principles (a "constitution") that were explicitly conceived to generate helpful and harmless behaviors. They dubbed this approach "Constitutional AI."[72]

To derive this initial constitution, the firm turned to a range of sources that included the UN Declaration of Human Rights, safety best practices from across the AI community, and Apple's terms of service.[73] In October of 2023, however, amidst a slew of more general concerns about AI governance, the firm launched a novel experiment in public governance, asking about 1,000 members of the American public to suggest the best rules for guiding a chatbot.[74] In describing this effort, the company acknowledged, "While constitutional AI is useful for making the normative values of our AI systems more transparent, it also highlights the outsized role we as developers play in selecting these values—after all, we wrote the constitution ourselves."[75]

Bringing the State Back In: Government Efforts to Regulate AI

Meanwhile, as firms were building their own hopes and channels for regulating AI, governments around the world were scrambling to develop more formal regulatory standards, well aware that the technology had already advanced far beyond the range of existing laws and standards. In both the United States and the European Union (EU), newly crafted laws such as the EU's General Data Protection Regulation (GDPR) and the state of California's Consumer Privacy Act (CCPA) gave end users greater control over their data, and outlined some basic provisions designed to minimize risks to individuals' privacy. But they offered few guidelines for tackling some of the more critical challenges posed by AI.

On June 14, 2023, the European Parliament passed a draft law of its AI Act, which, if enacted, would be the first AI-specific law passed by a major regulator.[76] The Act adopted a product-safety perspective, classifying AI use cases into unacceptable, high, limited, and minimal risk categories.[77] High risk systems would have to comply with a number of key requirements, including proper data governance, human oversight of the system, auditability, and security.[78] AI systems classified as unacceptable, such as those designed to deliberately manipulate users, were explicitly banned.[79] While advocates championed the AI Act as an important milestone that would have an immediate impact across the world, critics worried it could stifle innovation. Others faulted the proposed regulation for focusing on specific use cases and, therefore, potentially failing to govern new applications that would surely arise as the field developed.[80]

In the United States, two key norm-setting measures emerged in 2022: NIST (National Institute of Standards of Technology) published drafts of its AI Risk Management Framework and the White

House Office of Science and Technology Policy released its "Blueprint for an AI Bill of Rights."[81] The latter document addressed automated systems with the potential to affect the American public's rights, opportunities, or access to critical resources or services and sought, accordingly, to protect AI's end users from unsafe or ineffective automated decision-making, algorithmic discrimination, and data privacy risks. Under its provisions, organizations would be required to inform end users whenever automated systems were being used to make decisions and users would have the right to request explanations of how automated decisions were made and to opt out of automated decision-making in favor of humans wherever appropriate. [82] Yet because the "Blueprint" was not a piece of legislation or other instrument with regulatory effect, its provisions had no legal or technical force.[83]

Well aware of this regulatory gap, the White House invited leading U.S. tech companies, including OpenAI, Google, Microsoft, Amazon and Meta, to meet at the White House in the summer of 2023 to discuss the future of AI regulation. After a full day of meetings, the Biden-Harris administration reported that it had secured "voluntary" commitments from each of the invited firms to abide by certain standards. The firms pledged to share more information, for instance, between each other, the government, and researchers. They committed to creating "robust" tools for letting users know when they were seeing AI-generated content and to prioritize research on the social risks of AI. Once again, however, the White House shared few concrete details about what actions the companies would be expected to take, or how these standards would be enforced.[84]

By contrast, China moved quickly over the course of 2023 to implement some of the world's strictest regulations on generative AI. In August, Beijing issued 24 guidelines for regulating AI in accordance with the country's values.[85] These included measures requiring AI platform providers to register their services with the government, conduct security reviews before launching products, and include conspicuous labels (such as watermarks) on all AI-generated content.[86] Generative AI products like chatbots were expected not to generate content "inciting subversion of national sovereignty or the overturn of the socialist system," or "[advocate] terrorism or extremism, [promote] ethnic hatred and ethnic discrimination, violence and obscenity, as well as fake and harmful information."[87] Several agencies, including the Cyberspace Administration of China, were set to monitor tech firms' compliance with these new regulations, and empowered to conduct spot-checks as they saw fit.

Finally, action—or at least conversation—was well underway on the international stage. In May 2023, the G7 called for the development of international standards for AI, stressing the importance of a coordinated, harmonized effort. The group's leaders agreed to create a forum known as the "Hiroshima AI process" and to work towards keeping AI "in line with our shared democratic values."[88] Shortly afterwards, the U.N. Security Council held its own first meeting to discuss the risks of AI and subsequently called for a global watchdog to create, monitor, and enforce AI regulations. "No country will be untouched by A.I.," explained Britain's foreign secretary James Cleverly, "so we must involve and engage the widest coalition of international actors from all sectors."[89]

Back to the Future?

As Chat GPT neared the one-year anniversary of its launch, the field of AI continued to advance at an unprecedented pace. Between 2022 and 2023, the industry's market size grew from $142 billion to $208 billion, and was expected to reach $1.85 trillion by 2030.[90] Existential fears about the power of super-intelligence still collided—often in the same firm, or even the same person—with more optimistic visions of its cost-saving, lifesaving potential.

In the United States, the Biden administration issued an Executive Order in October 2023 to create AI safeguards, requiring that companies report to the federal government the risks their systems could pose to national security. To lessen the danger of Deep Fakes, particularly in relation to election interference, the order also encouraged tech firms to clearly mark their AI-generated content so that it would not be confused with the work of real people.[91] By contrast, just weeks before, Marc Andreesen, the co-creator of Netscape and one of Silicon Valley's leading venture investors, had published a 5,000-word "Techno-Optimist Manifesto" arguing that technology should not be feared but embraced as humanity's greatest tool for advancing society. "Technology," he wrote, "is the glory of human ambition and achievement, the spearhead for progress, and the realization of our potential." "Any deceleration of AI," he added, "will cost lives. Deaths that were preventable by the AI that was prevented from existing is a form of murder."[92]

The terms of this debate were highlighted in May of 2023, when Sam Altman and several other industry experts were called to testify before members of the U.S. Senate Subcommittee on Privacy, Technology, and the Law. One after another, senators from both parties stated their concerns about the implications of AI, and the desirability of putting some guardrails around it. Senator Josh Hawley (R-MO), for example, noted his fears of a coming "loss of jobs, invasion of privacy…manipulation of personal behavior, manipulation of personal opinions, and potentially the degradation of free elections in America."[93] Senator Cory Booker (D-NJ) asked, "Are you worried about the corporate concentration in this space and what effect it might have?… Are you saying that…my dreams of technology further democratizing opportunity and more are possible within a technology that is ultimately, I think, gonna be very centralized to a few players who already control so much?"[94] Conceptually, Altman appeared to agree with the lawmakers' concerns, acknowledging during his testimony that, "I think if this technology goes wrong, it can go quite wrong…We want to work with the government to prevent that from happening."[95] He reiterated similar views in a May 2024 speech at Harvard University, noting that, "[g]iven the magnitude of economic change, how we think about the illusion of the social contract…[and] how people like the balance between capital and labor, that's going to require serious societal debate...One of the most important things to figure out is how government can help play a role."[96]

But in practice, evidence of any self-inflicted guardrails was difficult to discern. On May 14, 2024, two of OpenAI's top executives resigned, with one posting on X that "I have been disagreeing with OpenAI leadership about the company's core priorities for quite some time, until we finally reached a breaking point."[97] Two weeks later, the company announced that it was training a new model that would soon bring AI to the "next level of capabilities" on the road to AGI. In the process, it also promised to create a new Safety and Security committee to help keep its technology safe.[98]

But did Altman and his competitors in the AI industry really have the ability or incentive to tackle the challenge of responsible AI? Or was it already too late?

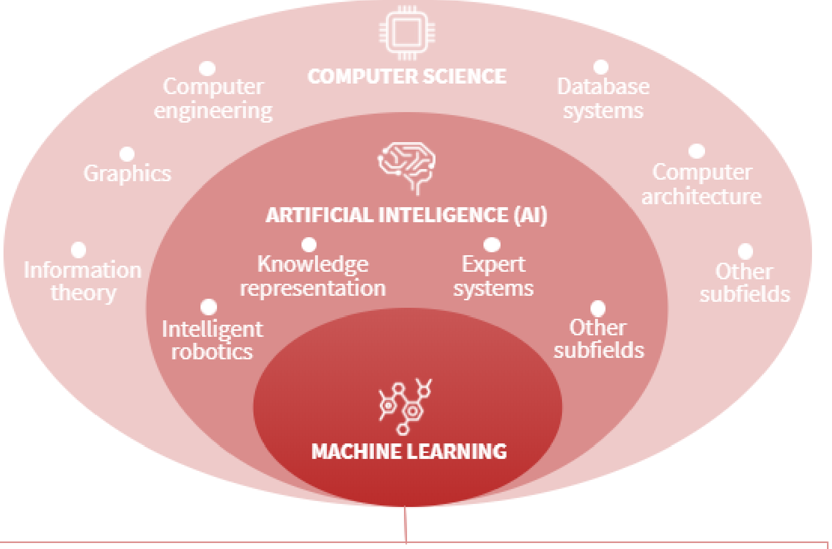
Exhibit 1 Taxonomy of AI and Machine Learning

ARTIFICIAL INTELIGENCE (AI)

is a scientific field within Computer Science, focusing on the study of computer systems that can perform tasks and solve problems that require human intelligence

MACHINE LEARNING

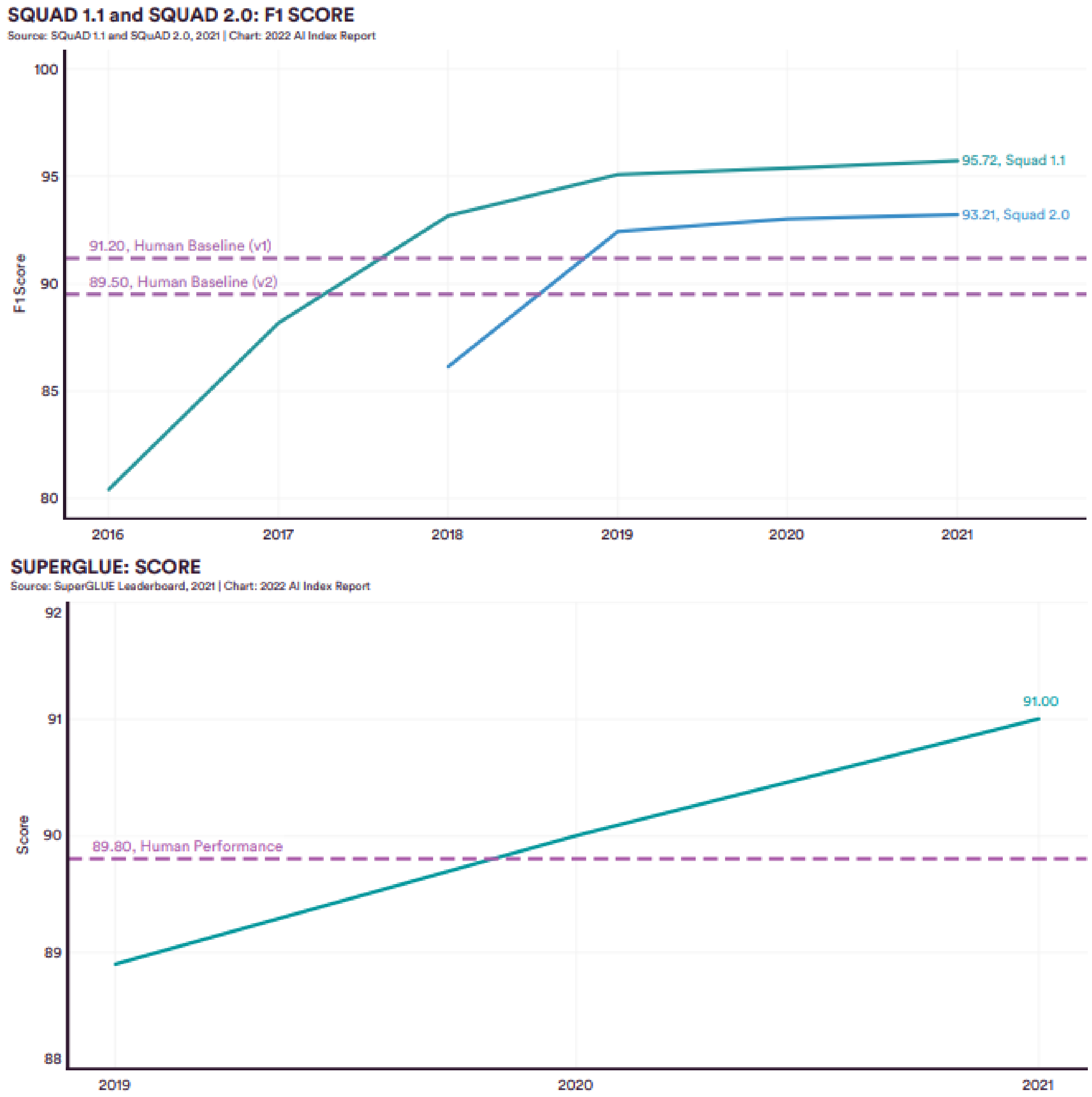
is a field within AI that focuses on a particular class of algorithms that can learn from data without being explicitly programmed



There are three main ways a machine can learn from data	**Supervised machine learning** Mapping inputs to labeled outputs	**Unsupervised machine learning** Finding patterns in unlabeled input data	**Reinforcement learning** Performing actions to maximize rewards
Each category of machine learning is effective in tackling particular kinds of tasks and problems	Commonly used in classification and regression problems such as: Natural language processing Image recognition Financial forecasting	Commonly used in segmentation and clustering problems such as: Pattern and trend recognition Customer segmentation Transaction monitoring	Similar to supervised learning, but reward mechanism in place instead of labelled output: Board and video games Robotics
There is a wide range of mathematical techniques that can be used to develop machine-learning models	Decision tree Random forest Gradient boosting Neural network Support vector machine Regularized regression	Clustering Dimensionality reduction Gaussian mixture model Principal component analysis Independent component analysis	Q-learning Temporal difference algorithm Neural network State-action-reward-state-action

Source: "The Risk of Machine-Learning Bias," Oliver Wyman, available at https://view.ceros.com/oliver-wyman/the-risk-of-machine-learning-bias/p/1, accessed May 2020.

Exhibit 2 AI Performance vs. Human Baseline on Reading Comprehension Tasks, 2016 to 2020



SQUAD 1.1 and SQUAD 2.0: F1 SCORE

Source: SQuAD 1.1 and SQuAD 2.0, 2021 | Chart: 2022 AI Index Report

SUPERGLUE: SCORE

Source: SuperGLUE Leaderboard, 2021 | Chart: 2022 AI Index Report

The Stanford Question Answering Dataset (SQuAD) benchmarks performance on reading comprehension. The dataset includes 107,785 question-and-answer pairs taken from 536 Wikipedia articles. Performance on SQuAD is measured by the F1 score, which is the average overlap between the AI system's answers and the actual correct answers: The higher the score, the better the performance.

Source: Daniel Zhang, Nestor Maslej, Erik Brynjolfsson, John Etchemendy, Terah Lyons, James Manyika, Helen Ngo, Juan Carlos Niebles, Michael Sellitto, Ellie Sakhaee, Yoav Shoham, Jack Clark, and Raymond Perrault, "The AI Index 2022 Annual Report," AI Index Steering Committee, Stanford Institute for Human-Centered AI, Stanford University, March 2022, pp. 75–76, https://stanford.io/3EP6kDT, accessed November 2022.

Exhibit 3 An Open Letter from Current and Former AI Company Employees

A Right to Warn about Advanced Artificial Intelligence

We are current and former employees at frontier AI companies, and we believe in the potential of AI technology to deliver unprecedented benefits to humanity.

We also understand the serious risks posed by these technologies. These risks range from the further entrenchment of existing inequalities, to manipulation and misinformation, to the loss of control of autonomous AI systems potentially resulting in human extinction. AI companies themselves have acknowledged these risks, as have governments across the world and other AI experts.

We are hopeful that these risks can be adequately mitigated with sufficient guidance from the scientific community, policymakers, and the public. However, AI companies have strong financial incentives to avoid effective oversight, and we do not believe bespoke structures of corporate governance are sufficient to change this.

AI companies possess substantial non-public information about the capabilities and limitations of their systems, the adequacy of their protective measures, and the risk levels of different kinds of harm. However, they currently have only weak obligations to share some of this information with governments, and none with civil society. We do not think they can all be relied upon to share it voluntarily.

So long as there is no effective government oversight of these corporations, current and former employees are among the few people who can hold them accountable to the public. Yet broad confidentiality agreements block us from voicing our concerns, except to the very companies that may be failing to address these issues. Ordinary whistleblower protections are insufficient because they focus on illegal activity, whereas many of the risks we are concerned about are not yet regulated. Some of us reasonably fear various forms of retaliation, given the history of such cases across the industry. We are not the first to encounter or speak about these issues.

We therefore call upon advanced AI companies to commit to these principles:

1. **That the company will not enter into or enforce** any agreement that prohibits "disparagement" or criticism of the company for risk-related concerns, nor retaliate for risk-related criticism by hindering any vested economic benefit;

2. **That the company will facilitate a verifiably anonymous process** for current and former employees to raise risk-related concerns to the company's board, to regulators, and to an appropriate independent organization with relevant expertise;

3. **That the company will support a culture of open criticism** and allow its current and former employees to raise risk-related concerns about its technologies to the public, to the company's board, to regulators, or to an appropriate independent organization with relevant expertise, so long as trade secrets and other intellectual property interests are appropriately protected;

4. **That the company will not retaliate against current and former employees who publicly share risk-related confidential information after other processes have failed.** We accept that any effort to report risk-related concerns should avoid releasing confidential information unnecessarily. Therefore, once an adequate process for anonymously raising concerns to the

> company's board, to regulators, and to an appropriate independent organization with relevant expertise exists, we accept that concerns should be raised through such a process initially. However, as long as such a process does not exist, current and former employees should retain their freedom to report their concerns to the public.

Signed by (alphabetical order):

Jacob Hilton, formerly OpenAI

Daniel Kokotajlo, formerly OpenAI

Ramana Kumar, formerly Google DeepMind

Neel Nanda, currently Google DeepMind, formerly Anthropic

William Saunders, formerly OpenAI

Carroll Wainwright, formerly OpenAI

Daniel Ziegler, formerly OpenAI

Anonymous, currently OpenAI

Anonymous, currently OpenAI

Anonymous, currently OpenAI

Anonymous, currently OpenAI

Anonymous, formerly OpenAI

Anonymous, formerly OpenAI

Endorsed by (in alphabetical order):

Yoshua Bengio

Geoffrey Hinton

Stuart Russell

June 4th, 2024

Source: "A Right to Warn About Advanced Artificial Intelligence," 4 June 2024, https://righttowarn.ai/, accessed June 2024.

Endnotes

[1] U.S. Atomic Energy Commission, *In the Matter of J. Robert Oppenheimer*, Transcript of Hearing Before Personnel Security Board, Washington D.C., 12 April 1954 through 6 May 1954.

[2] As quoted by his biographer, Walter Isaacson, at "Does Elon Musk Have Too Much Power?" Honestly with Bari Weiss, 7 November 2023, https://www.honestlypod.com/podcast/episode/2be70f0a/does-elon-musk-have-too-much-power, accessed November 2023.

[3] Rockwell Anyoha, "The History of Artificial Intelligence" Harvard University Graduate School of Arts and Sciences, https://sitn.hms.harvard.edu/flash/2017/history-artificial-intelligence/, accessed October 2023.

[4] Eric Horvitz, "Eric Horvitz: AI: Challenges, Advances, and Responsibilities," amphilsoc - American Philosophical Society, YouTube, June 6, 2022, https://www.youtube.com/watch?v=AUlco4XzqQs, accessed October 2022.

[5] Lenore Elle Hawkins, Chris Versace, and Mark Abssy, "What Is Artificial Intelligence and Who Are the Major Players in the Space?," Nasdaq, July 7, 2022, https://bit.ly/3gUNRh7, accessed October 2022.

[6] Alan Turing, "'Computing machinery and intelligence,'" *Mind* vol. 59 (October 1950), scanned by the A.M. Turing Trust from the estate of Mrs. Sara Turing, with the permission of Oxford University Press, available at https://bit.ly/3VvSjSX, accessed December 2022.

[7] "Microsoft using AI to make the world a better place: Satya Nadella," *The Economic Times,* 28 September 2023, https://economictimes.indiatimes.com/magazines/panache/microsoft-using-ai-to-make-the-world-a-better-place-satya-nadella/articleshow/54546010.cms, accessed November 2023.

[8] Bernard Marr, "How AI And Machine Learning Will Impact The Future Of Healthcare," Forbes, September 14, 2022, https://www.forbes.com/sites/bernardmarr/2022/09/14/how-ai-and-machine-learning-will-impact-the-future-of-healthcare/?sh=149d07c647e5, accessed November 2022. Also see E. Horvitz, Reflections on the Status and Future of Artificial Intelligence, Testimony Before the United States Senate, Hearing on the Dawn of Artificial Intelligence, Committee on Commerce Subcommittee on Space, Science, and Competitiveness, November 30, 2016, https://erichorvitz.com/Senate_Testimony_Eric_Horvitz.pdf; E. Horvitz. From Data to Predictions and Decisions: Enabling Evidence-Based Healthcare, Data Analytic Series, Computing Community Consortium, Computing Research Association (CRA), September 2010, https://erichorvitz.com/Evidence_based_healthcare_essay.pdf; E. Horvitz, The Future of Biomedical Informatics: Bottlenecks and Opportunities, In: E.H. Shortliffe, J.J. Cimino, et. al, Biomedical Informatics: Computer Applications in Health Care and Biomedicine, Springer, 2021, https://erichorvitz.com/Industry_research_perspective_on_biomedical_informatics.pdf.

[9] Jonathan Guo and Bin Li, "The Application of Medical Artificial Intelligence Technology in Rural Areas of Developing Countries," *Health Equity* 2(1):174-181, accessed November 2022.

[10] "Chase Taps Machine Learning For Proactive Approach To Fraud," PYMNTS, October 2, 2019, https://www.pymnts.com/fraud-prevention/2019/chase-taps-machine-learning-proactive-approach-fraud-strategy/, accessed November 2022; and Alyssa Schroer, "25 Examples of AI in Finance," BuiltIn, July 11, 2022, https://builtin.com/artificial-intelligence/ai-finance-banking-applications-companies, accessed November 2022.

[11] Lenore Elle Hawkins, Chris Versace, and Mark Abssy, "What Is Artificial Intelligence and Who Are the Major Players in the Space?," Nasdaq, July 7, 2022, https://bit.ly/3gUNRh7, accessed October 2022.

[12] Prableen Bajpai, "Microsoft (MSFT) and Artificial Intelligence," *Nasdaq*, August 7, 2019, https://www.nasdaq.com/articles/microsoft-msft-and-artificial-intelligence-2019-08-07; Lenore Elle Hawkins, Chris Versace, and Mark Abssy, "What Is Artificial Intelligence and Who Are the Major Players in the Space?," Nasdaq, July 7, 2022, https://bit.ly/3gUNRh7, accessed October 2022.

[13] Alan Turing, "Can Digital Computers Think?" *BBC Third Program*, May 15, 1951, scanned from the Turing Papers collection of the Archive Centre at King's College, Cambridge, assembled and made available by the A.M. Turing Trust, http://www.turingarchive.org/browse.php/B/5, accessed December 2022.

[14] Allan Defoe and Stuart Russell, "Yes, We Are Worried About the Existential Risk of Artificial Intelligence," *MIT Technology Review*, November 2, 2016, https://www.technologyreview.com/2016/11/02/156285/yes-we-are-worried-about-the-existential-risk-of-artificial-intelligence/, accessed December 2022.

[15] Alec Tyson and Emma Kikuchi, "Growing public concern about the role of artificial intelligence in daily life," Pew Research Center, 28 August 2023, https://www.pewresearch.org/short-reads/2023/08/28/growing-public-concern-about-the-role-of-artificial-intelligence-in-daily-life/, accessed November 2023.

[16] James Dillard, "Will AI wipe out humanity before the year 2100?" Manifold, https://manifold.markets/jamesdillard/will-ai-wipe-out-humanity-before-th, accessed November 2023.

[17] Kweilin Ellingrud, Saurabh Sanghvi, Gurneet Singh Dandona, Anu Madgavkar, Michael Chui, Olivia White, and Paige Hasebe, "Generative AI and the future of work in America," McKinsey Global Institute, 26 July 2023, https://www.mckinsey.com/mgi/our-research/generative-ai-and-the-future-of-work-in-america, accessed November 2023.

[18] Susan Lund and James Manyika, "Five lessons from history on AI, automation, and employment," McKinsey & Company, 28 November 2017, https://www.mckinsey.com/featured-insights/future-of-work/five-lessons-from-history-on-ai-automation-and-employment, accessed November 2023.

[19] Erik Brynjolfsson, "The Turing Trap: The Promise & Peril of Human-Like Artificial Intelligence," Stanford Digital Economy Lab, 12 January 2022, https://digitaleconomy.stanford.edu/news/the-turing-trap-the-promise-peril-of-human-like-artificial-intelligence/, accessed November 2023.

[20] Jane Wakefield, "AI emotion-detection software tested on Uyghurs," *BBC*, May 26, 2021, https://bbc.in/3igyzUo, accessed November 2022.

[21] Joy Buolamwini and Timnit Gebru, "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification," *Proceedings of Machine Learning Research* 81:1–15 (2018), eds. Sorelle A. Friedler and Christo Wilson, Conference on Fairness, Accountability, and Transparency, p. 1, https://bit.ly/3g3u9zn, accessed October 2022.

[22] Christina Pazzanese, "Great Promise But Potential for Peril," *The Harvard Gazette*, October 26 2020, https://news.harvard.edu/gazette/story/2020/10/ethical-concerns-mount-as-ai-takes-bigger-decision-making-role/, accessed December 2022.

[23] See, for example, Emmanuel Martinez and Lauren Kirchner, "The secret bias hidden in mortgage-approval algorithms," Associated Press, 25 August 2021, https://apnews.com/article/lifestyle-technology-business-race-and-ethnicity-mortgages-2d3d40d5751f933a88c1e17063657586, accessed November 2023; and Pranshu Verma, "These robots were trained on AI. They became racist and sexist," *The Washington Post,* 16 July 2022, https://www.washingtonpost.com/technology/2022/07/16/racist-robots-ai/, accessed November 2023; and Miranda Bogen, "All the Ways Hiring Algorithms Can Introduce Bias," *Harvard Business Review*, 6 May 2019, https://hbr.org/2019/05/all-the-ways-hiring-algorithms-can-introduce-bias, accessed November 2023.

[24] Julia Angwin, Jeff Larsen, Surya Mattu, and Lauren Kirchna, "Machine Bias", Pro Publica, May 23, 2016, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing, accessed December 2022.

[25] Julia Angwin, Jeff Larsen, Surya Mattu, and Lauren Kirchna, "Machine Bias", Pro Publica, May 23, 2016, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing, accessed December 2022.

[26] R. H. Coase, "The Problem of Social Cost," Journal of Law and Economics 3 (1960), p. 6, available via https://www2.econ.iastate.edu/classes/tsc220/hallam/Coase.pdf, accessed June 2024.

[27] Friedman was insistent on this point. See Milton Friedman, *Capitalism and Freedom* (Chicago: University of Chicago, 1962).

[28] Amy R. Poteete, Marco A. Janssen, and Elinor Ostrom, *Working Together: Collective Action, the Commons, and Multiple Methods in Practice*, (Princeton and Oxford: Princeton University Press, 2010), p. 215-217.

[29] "Model Rules of Professional Conduct," American Bar Association, https://www.americanbar.org/groups/professional_responsibility/publications/model_rules_of_professional_conduct/ , accessed June 2024; and "Code of Medical Ethics," American Medical Association, https://code-medical-ethics.ama-assn.org/, accessed June 2024.

[30] "Vision and Mission," Bluetooth, https://www.bluetooth.com/about-us/vision/, accessed June 2024; and "About 3GPP," 3rd Generation Partnership Project, https://www.3gpp.org/about-us, accessed June 2024.

[31] Adonis Hoffman, "Why self-regulation is best for artificial intelligence," *The Hill*, 8 November 2023, https://thehill.com/opinion/4300288-why-self-regulation-is-best-for-artificial-intelligence/, accessed June 2024; and see Michael A. Cusumano, Annabelle Gawer, and David B. Yoffie, "Social Media Companies Should Self-Regulate. Now." *Harvard Business Review*, 15 January 2021, https://hbr.org/2021/01/social-media-companies-should-self-regulate-now, accessed June 2024.

[32] Virginia Haufler, *A Public Role for the Private Sector: Industry Self-Regulation In a Global Economy* (Washington, DC: Carnegie Endowment for International Peace, 2001), p. 2

[33] Karen Hao, "The messy, secretive reality behind OpenAI's bid to save the world," MIT Technology Review, 17 February 2020, https://www.technologyreview.com/2020/02/17/844721/ai-openai-moonshot-elon-musk-sam-altman-greg-brockman-messy-secretive-reality/, accessed October 2023.

[34] "Introducing OpenAI," OpenAI, https://openai.com/blog/introducing-openai, accessed October 2023.

[35] Karen Hao, "The messy, secretive reality behind OpenAI's bid to save the world," MIT Technology Review, 17 February 2020, https://www.technologyreview.com/2020/02/17/844721/ai-openai-moonshot-elon-musk-sam-altman-greg-brockman-messy-secretive-reality/, accessed October 2023.

[36] "OpenAI Charter," OpenAI, 9 April 2018, https://openai.com/charter, accessed October 2023.

[37] Reed Albergotti, "The secret history of Elon Musk, Sam Altman, and OpenAI," Semafor, 24 March 2023, https://www.semafor.com/article/03/24/2023/the-secret-history-of-elon-musk-sam-altman-and-openai, accessed October 2023.

[38] James Vincent, "Elon Musk leaves board of AI safety group to avoid conflict of interest with Tesla," The Verge, 21 February 2018, https://www.theverge.com/2018/2/21/17036214/elon-musk-openai-ai-safety-leaves-board, accessed October 2023.

[39] Karen Hao, "The messy, secretive reality behind OpenAI's bid to save the world," MIT Technology Review, 17 February 2020, https://www.technologyreview.com/2020/02/17/844721/ai-openai-moonshot-elon-musk-sam-altman-greg-brockman-messy-secretive-reality/, accessed October 2023.

[40] Reed Albergotti, "The secret history of Elon Musk, Sam Altman, and OpenAI," Semafor, 24 March 2023, https://www.semafor.com/article/03/24/2023/the-secret-history-of-elon-musk-sam-altman-and-openai, accessed October 2023.

[41] "DALL-E: Creating images from text," OpenAI, 5 January 2021, https://openai.com/research/dall-e, accessed November 2023.

[42] Aleksandra Yosifova, "The Evolution of ChatGPT: History and Future," 365 Data Science, 14 August 2023, https://365datascience.com/trending/the-evolution-of-chatgpt-history-and-future/, accessed October 2023.

[43] Kevin Roose, "How ChatGPT Kicked Off an A.I. Arms Race," *New York Times,* 3 February 2023, https://www.nytimes.com/2023/02/03/technology/chatgpt-openai-artificial-intelligence.html, accessed October 2023.

[44] Grace Kay, "The history of ChatGPT creator OpenAI, which Elon Musk helped found before parting ways and criticizing," Business Insider, 1 February 2023, https://www.businessinsider.com/history-of-openai-company-chatgpt-elon-musk-founded-2022-12, accessed October 2023.

[45] Nico Grant and Cade Metz, "A New Chat Bot Is a 'Code Red' for Google's Search Business," *New York Times,* 21 December 2023, https://www.nytimes.com/2022/12/21/technology/ai-chatgpt-google-search.html, accessed November 2023; and Andrew R. Chow and Billy Perrigo, "The AI Arms Race is Changing Everything," *TIME*, 17 February 2023, https://time.com/6255952/ai-impact-chatgpt-microsoft-google/, accessed November 2023.

[46] Coco Feng, Chinese firms scramble to take advantage of ChatGPT concept, even though the service is not officially available," *South China Morning Post,* 7 February 2023, https://www.scmp.com/tech/article/3209359/chinese-firms-scramble-take-advantage-chatgpt-concept-even-though-service-not-officially-available, accessed November 2023.

[47] Andrew R. chow and Billy Perrigo, "The AI Arms Race is Changing Everything," *TIME*, 17 February 2023, https://time.com/6255952/ai-impact-chatgpt-microsoft-google/, accessed November 2023.

[48] Satya Nadella, "The Partnership of the Future," SLATE, 28 June 2016, https://slate.com/technology/2016/06/microsoft-ceo-satya-nadella-humans-and-a-i-can-work-together-to-solve-societys-challenges.html, accessed October 2023; and "Facebook's five pillars of Responsible AI," Meta, 22 June 2021, https://ai.meta.com/blog/facebooks-five-pillars-of-responsible-ai/, accessed October 2023; and Peter Hallinan, "AWS Reaffirms its Commitment to Responsible Generative AI," Amazon, 25 July 2023, https://aws.amazon.com/blogs/machine-learning/aws-reaffirms-its-commitment-to-responsible-generative-ai/, accessed October 2023; and "Our Principles," Google AI, https://ai.google/responsibility/principles/, accessed October 2023.

[49] Nathan Reiff, "10 Biggest Software Companies," Investopedia, 24 May 2023, https://www.investopedia.com/articles/personal-finance/121714/worlds-top-10-software-companies.asp#toc-1-microsoft-corp-msft, accessed November 2023.

[50] Microsoft, "Microsoft expands artificial intelligence (AI) efforts with creation of new Microsoft AI and Research Group," September 29, 2016, https://news.microsoft.com/2016/09/29/microsoft-expands-artificial-intelligence-ai-efforts-with-creation-of-new-microsoft-ai-and-research-group/, accessed October 2022.

[51] Microsoft, "Microsoft announces new tools and services to help developers modernize existing apps and build more intelligent ones, on every platform, for every platform," May 10, 2017, https://news.microsoft.com/2017/05/10/microsoft-announces-new-tools-services-help-developers-modernize-existing-apps-build-intelligent-ones-every-platform-every-platform/, accessed October 2022.

[52] "Build an AI-powered organization," Microsoft, https://www.microsoft.com/en-us/ai/industry/ai-in-business, accessed October 2022.

[53] "Build an AI-powered organization," Microsoft, https://www.microsoft.com/en-us/ai/industry/ai-in-business, accessed October 2022.

[54] "Microsoft Responsible AI Standard, v2: General Requirements," Microsoft, June 2022, p.3, https://bit.ly/3CUWBuU, accessed October 2022.

[55] "Google achieves AI 'breakthrough' by beating Go champion," BBC News, 27 January 2023, https://www.bbc.com/news/technology-35420579, accessed November 2023.

[56] Gintaras Radauskas, "Google will protect users in AI copyright accusations, Bard excluded," CyberNews, 13 October 2023, https://cybernews.com/news/google-cloud-ai-copyright-infringement-legal/, accessed October 2023.

[57] Sundar Pichai, "Google CEO: Building AI responsibly is the only race that really matters," *Financial Times*, 23 May 2023, https://www.ft.com/content/8be1a975-e5e0-417d-af51-78af17ef4b79, accessed October 2023.

[58] Eamon Javers, "How Google's former CEO Eric Schmidt helped write A.I. laws in Washington without publicly disclosing investments in A.I. startups," CNBC, 24 October 2022, https://www.cnbc.com/2022/10/24/how-googles-former-ceo-eric-schmidt-helped-write-ai-laws-in-washington-without-publicly-disclosing-investments-in-ai-start-ups.html, accessed November 2023.

[59] Will Douglas Heaven, "Why Meta's latest large language model survived only three days online," MIT Technology Review, 18 November 2023, https://www.technologyreview.com/2022/11/18/1063487/meta-large-language-model-ai-only-survived-three-days-gpt-3-science/, accessed November 2023.

[60] Will Douglas Heaven, "Why Meta's large language model survived only three days online," MIT Technology Review, 18 November 2022, https://www.technologyreview.com/2022/11/18/1063487/meta-large-language-model-ai-only-survived-three-days-gpt-3-science/, accessed November 2023.

[61] Mike Isaac and Cade Metz, "Meet the A.I. Jane Austen: Meta Weaves A.I. Throughout Its Apps," *New York Times,* 27 September 2023, https://www.nytimes.com/2023/09/27/technology/meta-ai-celebrities.html, accessed October 2023.

[62] Max Chafkin, "Meta's New AI Buddies Aren't Great Conversationalists," *Bloomberg*, 17 October 2023, https://www.bloomberg.com/news/newsletters/2023-10-17/meta-s-celebrity-ai-chatbots-on-facebook-instagram-are-surreal, accessed November 2023.

[63] Andrew Tarantola, "Meta AI is a new chatbot platform that starts with over 25 personalities," Engadget, 27 September 2023, https://www.engadget.com/meta-is-unleashing-more-than-25-ai-chatbot-personalities-onto-the-world-181428710.html, accessed November 2023.

[64] Mike Isaac and Cade Metz, "Meet the A.I. Jane Austen: Meta Weaves A.I. Throughout Its Apps," *New York Times,* 27 September 2023, https://www.nytimes.com/2023/09/27/technology/meta-ai-celebrities.html, accessed October 2023; and Aleksandra Yosifova, "The Evolution of ChatGPT: History and Future," 365 Data Science, 14 August 2023, https://365datascience.com/trending/the-evolution-of-chatgpt-history-and-future/, accessed October 2023.

[65] "Building Generative AI Features Responsibly," Meta, 27 September 2023, https://about.fb.com/news/2023/09/building-generative-ai-features-responsibly/, accessed October 2023.

[66] "Building Generative AI Features Responsibly," Meta, 27 September 2023, https://about.fb.com/news/2023/09/building-generative-ai-features-responsibly/, accessed October 2023.

[67] Edd Gent, "Protestors Decry Meta's 'Irreversible Proliferation' of AI," IEEE Spectrum, 6 October 2023, https://spectrum.ieee.org/meta-ai, accessed November 2023.

[68] Cade Metz and Mike Isaac, "In Battle Over A.I., Meta Decides to Give Away Its Crown Jewels," *New York Times,* 18 May 2023, https://www.nytimes.com/2023/05/18/technology/ai-meta-open-source.html, accessed December 2023.

[69] Edd Gent, "Protestors Decry Meta's 'Irreversible Proliferation' of AI," IEEE Spectrum, 6 October 2023, https://spectrum.ieee.org/meta-ai, accessed November 2023.

[70] Will Henshall and Billy Perrigo, "Dario and Daniela Amodei: Time 100 AI" *Time*, 7 September 2023, https://time.com/collection/time100-ai/6309047/daniela-and-dario-amodei/, accessed May 2024.

[71] "The Long-Term Benefit Trust," Anthropic, 19 September 2023, https://www.anthropic.com/index/the-long-term-benefit-trust, accessed December 2023.

[72] "Constitutional AI: Harmlessness from AI Feedback," Anthropic, https://www-files.anthropic.com/production/images/Anthropic_ConstitutionalAI_v2.pdf?dm=1694134767, accessed December 2023.

[73] "Claude's Constitution," Anthropic, 9 May 2023, https://www.anthropic.com/index/claudes-constitution, accessed December 2023.

[74] "Collective Constitutional AI: Aligning a Language Model with Public Input," Anthropic, 17 October 2023, https://www.anthropic.com/index/collective-constitutional-ai-aligning-a-language-model-with-public-input, accessed December 2023.

[75] "Collective Constitutional AI: Aligning a Language Model with Public Input," Anthropic, 17 October 2023, https://www.anthropic.com/index/collective-constitutional-ai-aligning-a-language-model-with-public-input, accessed December 2023.

[76] Adam Satariano, "E.U. Takes Major Step Toward Regulating A.I.," *New York Times*, 14 June 2023, https://www.nytimes.com/2023/06/14/technology/europe-ai-regulation.html, accessed October 2023; and European Commission, Shaping Europe's Digital Future, "Proposal for a Regulation laying down harmonised rules on artificial intelligence," Policy and Legislation, April 21, 2021, https://bit.ly/3FXaRWU, accessed December 2022.

[77] Eve Gaumond, "Artificial Intelligence Act: What Is the European Approach for AI?," *Lawfare*, June 4, 2021, https://bit.ly/3htWJe5, accessed December 2022.

[78] Eve Gaumond, "Artificial Intelligence Act: What Is the European Approach for AI?," *Lawfare*, June 4, 2021, https://bit.ly/3htWJe5, accessed December 2022.

[79] Eve Gaumond, "Artificial Intelligence Act: What Is the European Approach for AI?," *Lawfare*, June 4, 2021, https://bit.ly/3htWJe5, accessed December 2022.

[80] Eve Gaumond, "Artificial Intelligence Act: What Is the European Approach for AI?," *Lawfare*, June 4, 2021, https://bit.ly/3htWJe5, accessed December 2022.

[81] The White House, "Blueprint for an AI Bill of Rights," Office of Science and Technology Policy, https://bit.ly/3W06Ar9, accessed December 2022.

[82] Alex Engler, "The AI Bill of Rights makes uneven progress on algorithmic protections," Brookings Institution, November 21, 2022, https://brook.gs/3Fp6Fxi, accessed December 2022.

[83] Kay Firth-Butterfield, Karen Silverman, and Benjamin Larsen, "Understanding the US 'AI Bill of Rights' - and how it can help keep AI Accountable," World Economic Forum, October 14, 2022, https://bit.ly/3HAL6gd, accessed December 2022.

[84] Robert Hart, "Tech Giants Make 'Voluntary' Pledge To Develop Responsible AI—Including OpenAI And Google—White House Says," Forbes, 21 July 2023, https://www.forbes.com/sites/roberthart/2023/07/21/tech-giants-make-voluntary-pledge-to-develop-responsible-ai-including-openai-and-google-white-house-says/?sh=482af368d33a, accessed October 2023.

[85] Will Henshall, "How China's New AI Rules Could Affect U.S. Companies," *Time,* 19 September 2023, https://time.com/6314790/china-ai-regulation-us/, accessed October 2023.

[86] Sarah Zheng and Jane Zhang, "China Tries to Balance State Control and State Support of AI," *TIME*, 14 August 2023, https://time.com/6304831/china-ai-regulations/, accessed October 2023.

[87] Will Henshall, "How China's New AI Rules Could Affect U.S. Companies," *Time,* 19 September 2023, https://time.com/6314790/china-ai-regulation-us/, accessed October 2023.

[88] Kantaro Komiya and Supantha Mukherjee, "G7 calls for developing global technical standards for AI," Reuters, 20 May 2023, https://www.reuters.com/world/g7-calls-developing-global-technical-standards-ai-2023-05-20/, accessed October 2023.

[89] Farnaz Fassihi, "U.N. Officials Urge Regulation of Artificial Intelligence," *New York Times*, 18 July 2023, https://www.nytimes.com/2023/07/18/world/un-security-council-ai.html, accessed October 2023.

[90] "Artificial intelligence (AI) market size worldwide in 2021 with a forecast until 2030," Statista, 2023, https://www.statista.com/statistics/1365145/artificial-intelligence-market-size/, accessed November 2023.

[91] Cecilia Kang and David Sanger, "Biden Issues Executive Order to Create A.I. Safeguards," *New York Times,* 30 October 2023, https://www.nytimes.com/2023/10/30/us/politics/biden-ai-regulation.html, accessed November 2023.

[92] Marc Aandreessen, "The Techno-Optimist Manifesto," Andreessen Horowitz, 16 October 2023, https://a16z.com/the-techno-optimist-manifesto/, accessed November 2023.

[93] "Transcript: Senate Judiciary Subcommittee Hearing on Oversight of AI," Tech Policy Press, 16 May 2023, https://www.techpolicy.press/transcript-senate-judiciary-subcommittee-hearing-on-oversight-of-ai/, accessed May 2024.

[94] "Transcript: Senate Judiciary Subcommittee Hearing on Oversight of AI," Tech Policy Press, 16 May 2023, https://www.techpolicy.press/transcript-senate-judiciary-subcommittee-hearing-on-oversight-of-ai/, accessed May 2024.

[95] Cecilia Kang, "OpenAI's Sam Altman Urges A.I. Regulation in Senate Hearing," *New York Times*, 16 May 2023, https://www.nytimes.com/2023/05/16/technology/openai-altman-artificial-intelligence-regulation.html, accessed November 2023.

[96] Ellen Chang, "'The Shape of Jobs Will Change,' Sam Altman Says at Harvard, Harvard Business School Institute for Business in Global Society, 1 May 2024, https://www.hbs.edu/bigs/sam-altman-harvard-business-school, accessed May 2024.

[97] Quoted in Sigal Samuel, "'I Lost Trust': Why the OpenAI team in charge of safeguarding humanity imploded," Vox, May 18, 2024, https://www.vox.com/future-perfect/2024/5/17/24158403/openai-resignations-ai-safety-ilya-sutskever-jan-leike-artificial-intelligence, accessed May 2024.

[98] Cade Metz, "OpenAI Says It Has Begun Training a New Flagship A.I. Model," *New York Times*, 28 May 2024, https://www.nytimes.com/2024/05/28/technology/openai-gpt4-new-model.html, accessed May 2024.