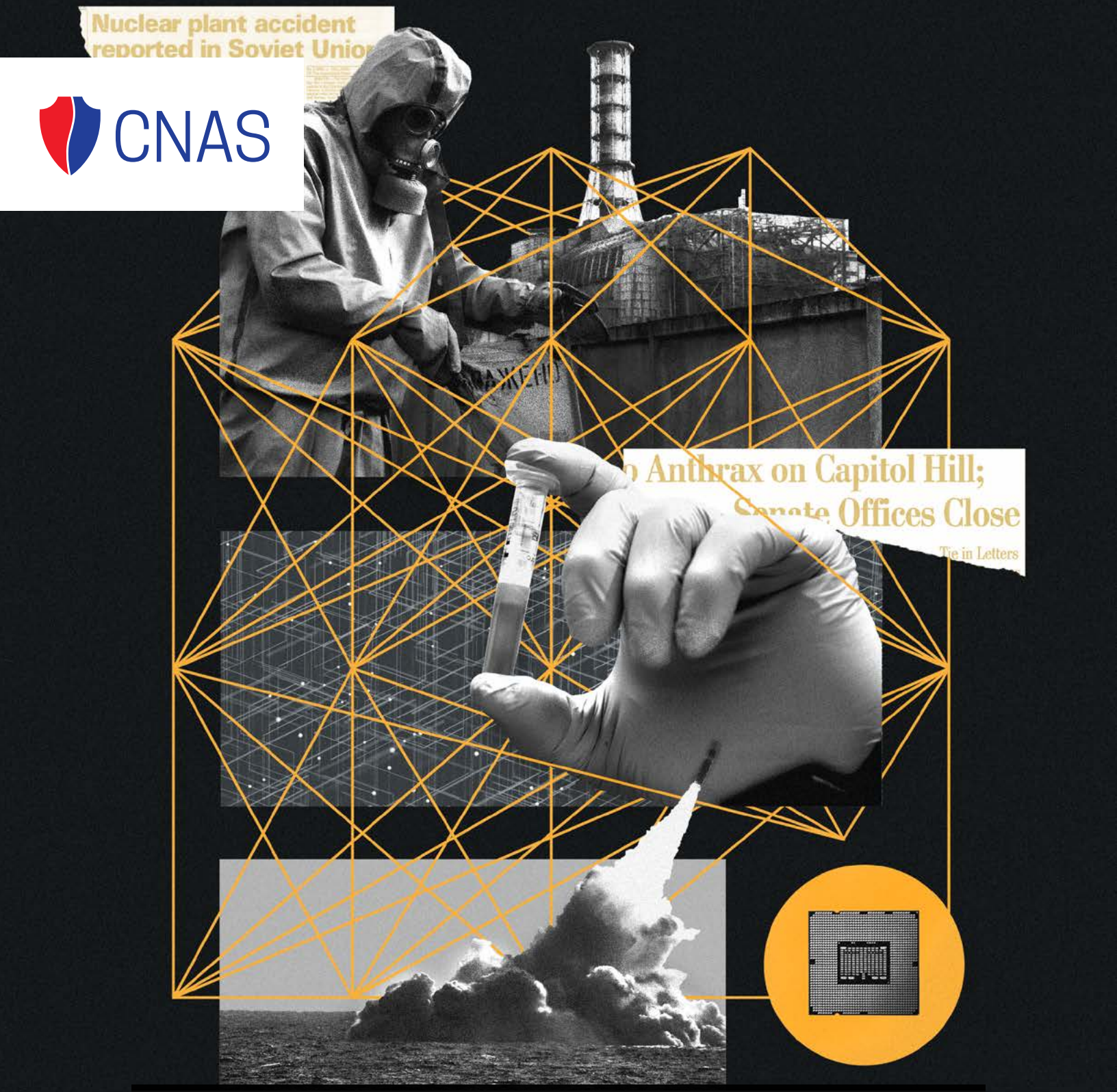


JUNE 2024

Catalyzing Crisis

A Primer on Artificial Intelligence, Catastrophes, and National Security

Bill Drexel and Caleb Withers

About the Authors

Bill Drexel is a fellow for the Technology and National Security Program at the Center for a New American Security (CNAS). His work focuses on Sino-American competition, artificial intelligence, and technology as an element of American grand strategy. Previously, Drexel worked on humanitarian innovation at the UN (International Organization for Migration) and on Indo-Pacific affairs at the American Enterprise Institute. Always seeking on-the-ground exposure, Drexel has served as a rescue boat driver during Libya's migration crisis; conducted investigative research in the surveillance state of Xinjiang, China; and supported humanitarian data efforts across wartime Ukraine. He holds a BA from Yale University and master's degrees from Cambridge and Tsinghua universities.

Caleb Withers is a research assistant for the Technology and National Security Program at CNAS. Before CNAS, he worked as a policy analyst for a variety of New Zealand government departments. He holds an MA in security studies from Georgetown University with a concentration in technology and security, and a bachelor's of commerce from Victoria University of Wellington with majors in economics and information systems.

Acknowledgments

The authors are grateful to Suzanne Spaulding and Andrew Imbrie for their valuable feedback and suggestions on earlier drafts of this report. This report would not have been possible without contributions from our CNAS colleagues, including Paul Scharre, Melody Cook, Rin Rothback, Allison Francis, Jake Penders, Tim Fist, Josh Wallin, Michael Depp, and Noah Greene. The report was made possible with the generous support of Open Philanthropy.

About the Technology & National Security Program

The CNAS Technology and National Security Program explores the policy challenges associated with emerging technologies. A key focus of the program is bringing together the technology and policy communities to better understand these challenges and together develop solutions.

About the Artificial Intelligence Safety & Stability Project

The CNAS AI Safety & Stability Project is a multiyear, multiprogram effort that addresses the established and emerging risks associated with artificial intelligence. The work is focused on anticipating and mitigating catastrophic AI failures, improving the U.S. Department of Defense's processes for AI testing and evaluation, understanding and shaping opportunities for compute governance, understanding Chinese decision-making on AI and stability, and understanding Russian decision-making on AI and stability.

TABLE OF CONTENTS

Executive Summary

The arrival of ChatGPT in November 2022 initiated both great excitement and fear around the world about the potential and risks of artificial intelligence (AI). In response, several AI labs, national governments, and international bodies have launched new research and policy efforts to mitigate large-scale AI risks. However, growing efforts to mitigate these risks have also produced a divisive and often confusing debate about how to define, distinguish, and prioritize severe AI hazards. This categorical confusion could complicate policymakers' efforts to discern the unique features and national security implications of the threats AI poses—and hinder efforts to address them. Specifically, emerging catastrophic risks with weighty national security implications are often overlooked between the two dominant discussions about AI concern in public discourse: present-day systemic harms from AI related to bias and discrimination on the one hand, and cantankerous, future-oriented debates about existential risks from AI on the other.

This report aims to:

Demonstrate the growing importance of mitigating AI's catastrophic risks for national security practitioners

Clarify what AI's catastrophic risks are (and are not)

Introduce the dimensions of AI safety that will most shape catastrophic risks

Catastrophic AI risks, like all catastrophic risks, demand attention from the national security community as a critical threat to the nation's health, security, and economy. In scientifically advanced societies like the United States, powerful technologies can pose outsized risks for catastrophes, especially in cases such as AI, where the technology is novel, fast-moving, and relatively untested. Given the wide range of potential applications for AI, including in biosecurity, military systems, and other high-risk domains, prudence demands proactive efforts to distinguish, prioritize, and mitigate risks. Indeed, past incidents related to finance, biological and chemical weapons, cybersecurity, and nuclear command and control all hint at possible AI-related catastrophes in the future, including AI-accelerated biological weapons of mass destruction (WMD) production, financial meltdowns from AI trading, or even accidental weapons exchanges from AI-enabled command and control systems. In addition to helping initiate crises, AI tools can also erode states' abilities to cope with them by degrading their public information ecosystems, potentially making catastrophes more likely and their effects more severe.

Perhaps the most confusing aspect of public discourse about AI risks is the inconsistent and sometimes interchangeable use of the terms "catastrophic risks" and "existential risks"—the latter often provoking strong disagreements among experts. To disentangle these concepts, it is helpful to consider different crises along a spectrum of magnitude, in which the relative ability of a state to respond to a crisis determines its classification. By this definition, a catastrophic event is one that requires the highest levels of state response, with effects that are initially unmanageable or mismanaged—causing large-scale losses of life or economic vitality. Existential risks are even larger in magnitude, threatening to overwhelm all states' ability to respond, resulting in the irreversible collapse of human civilization or the extinction of humanity. Both differ from smaller-scale crises, such as emergencies and disasters, which initiate local and regional state crisis management responses, respectively. While the prospect of existential risks unsurprisingly provokes pitched disagreements and significant media attention, catastrophic risks are of nearer-term relevance, especially to national security professionals. Not only are catastrophic risks less speculative, but the capabilities that could enable AI catastrophes are also closer to development than those that would be of concern for existential risks. Catastrophic AI risks are also, in many cases, variants on issues that the U.S. government has already identified as high priorities for national security, including possibilities of nuclear escalation, biological attacks, or financial meltdowns.

Despite recent public alarm concerning the catastrophic risks of powerful "deep learning"–based AI tools in particular, the technology's integration into high-risk domains is largely still in its nascent forms, giving the U.S. government and industry the opportunity to help develop the technology with risk mitigation in mind. But accurately predicting the full range of the most likely AI catastrophes and their impacts is challenging for several reasons, particularly as emerging risks will depend on the ways in which AI tools are integrated into high-impact domains with the potential to disrupt society. Instead, this report distills prior research across a range of fields into four dimensions of AI safety shaping AI's catastrophic risks. Within each dimension, the report outlines each issue's dynamics and relevance to catastrophic risk.

Safety Dimension	Question	Issues	
New capabilities	What dangers arise from new AI-enabled capabilities across different domains?	▪ Dangerous capabilities ▪ Emergent capabilities ▪ Latent capabilities	
Technical safety challenges	In what ways can technical failures in AI-enabled systems escalate risks?	▪ Alignment, specification gaming ▪ Loss of control ▪ Robustness	▪ Calibration ▪ Adversarial attacks ▪ Explainability and interpretability
Integrating AI into complex systems	How can the integration of AI into high-risk systems disrupt or derail their operations?	▪ Automation bias ▪ Operator trust ▪ The lumberjack effect ▪ Eroded sensitivity to operations	▪ Deskilling, enfeeblement ▪ Tight coupling ▪ Emergent behavior ▪ Release and proliferation
Conditions of AI development	How do the conditions under which AI tools are developed influence their safety?	▪ Corporate and geopolitical competitive pressures ▪ Deficient safety cultures ▪ Systemic underinvestment in technical safety R&D	▪ Social resilience ▪ Engineering memory life cycles

Though presented individually, in practice the issues described are most likely to lead to catastrophic outcomes when they occur in combination. Taken together, perhaps the most underappreciated feature of emerging catastrophic AI risks from this exploration is the outsized likelihood of AI catastrophes originating from China. There, a combination of the Chinese Communist Party's efforts to accelerate AI development, its track record of authoritarian crisis mismanagement, and its censorship of information on accidents all make catastrophic risks related to AI more acute.

To address emerging catastrophic risks associated with AI, this report proposes that:

- AI companies, government officials, and journalists should be more precise and deliberate in their use of terms around AI risks, particularly in reference to "catastrophic risks" and "existential risks," clearly differentiating the two.
- Building on the Biden administration's 2023 executive order on AI, the departments of Defense, State, Homeland Security, and other relevant government agencies should more holistically explore the risks of AI integration into high-impact domains such as biosecurity, cybersecurity, finance, nuclear command and control, critical infrastructure, and other high-risk industries.
- Policymakers should support enhanced development of testing and evaluation for foundation models' capabilities.
- The U.S. government should plan for AI-related catastrophes abroad that might impact the United States, and mitigate those risks by bolstering American resilience.
- The United States and allies must proactively establish catastrophe mitigation measures internationally where appropriate, for example by building on their promotion of responsible norms in autonomous weapons and AI in nuclear command.

AI-related catastrophic risks may seem complex and daunting, but they remain manageable. While national security practitioners must appraise these risks soberly, they must also resist the temptation to over-fixate on worst-case scenarios at the expense of pioneering a strategically indispensable, powerful new technology. To this end, efforts to ensure robust national resilience against AI's catastrophic risks go hand in hand with pursuing the immense benefits of AI for American security and competitiveness.

Introduction

Since ChatGPT was launched in November 2022, artificial intelligence (AI) systems have captured public imagination across the globe. ChatGPT's record-breaking speed of adoption—logging 100 million users in just two months—gave an unprecedented number of individuals direct, tangible experience with the capabilities of today's state-of-the-art AI systems.[1] More than any other AI system to date, ChatGPT and subsequent competitor large language models (LLMs) have awakened societies to the promise of AI technologies to revolutionize industries, cultures, and political life. This public recognition follows from a growing awareness in the U.S. government that AI, in the words of the National Security Commission on Artificial Intelligence, "will be the most powerful tool in generations for benefiting humanity," and an indispensable strategic priority for continued American leadership.[2]

But alongside the excitement surrounding ChatGPT is growing alarm about myriad risks from emerging AI capabilities. These range from systemic bias and discrimination to labor automation, novel biological and chemical weapons, and even—some experts argue—the possibility of human extinction. The sudden explosion of attention to such diverse concerns has ignited fierce debates about how to characterize and prioritize such risks. Leading AI labs and policymakers alike are beginning to devote considerable attention to catastrophic risks stemming from AI specifically: OpenAI launched a purpose-built Preparedness team to address these risks, just as Anthropic crafted a Responsible Scaling Policy to "require safety, security, and operational standards appropriate to a model's potential for catastrophic risk."[3] In November 2023, 28 countries signed the Bletchley Declaration, a statement resulting from the United Kingdom's (UK's) AI Safety Summit, that likewise affirmed AI's potential to produce "catastrophic" harms.[4]

For national security practitioners, the maelstrom of often-conflicting opinions about the potential harms of AI can obscure emerging catastrophic risks with direct national security implications. Between the attention devoted to the range of harms AI is already causing in bias, discrimination, and systemic impacts on the one hand, and the focus on future-oriented debates about existential risks posed by AI on the other, these emerging catastrophic threats can be easily overlooked. That would be a major mistake: progress in AI could enable or contribute to scenarios that have debilitating effects on the United States, from enhanced bioterrorism to nationwide financial meltdowns to unintended nuclear exchanges. Given the potential magnitude of these events, policymakers urgently need sober analysis to better understand the emerging risks of AI-enabled catastrophes. Better clarity about the large-scale risks of AI need not inhibit the United States' competitiveness in developing this strategically indispensable technology in the years ahead, as some fear. To the contrary, a more robust understanding of large-scale risks related to AI may help the United States to forge ahead with greater confidence, and to avoid incidents that could hamstring development due to public backlash.

This report aims to help policymakers understand catastrophic AI risks and their relevance to national security in three ways. First, it attempts to further clarify AI's catastrophic risks and distinguish them from other threats such as existential risks that have featured prominently in public discourse. Second, the report explains why catastrophic risks associated with AI development merit close attention from U.S. national security practitioners in the years ahead. Finally, it presents a framework of AI safety dimensions that contribute to catastrophic risks.

Despite recent public alarm concerning the catastrophic risks of AI, the technology's integration into high-risk domains is largely still in its nascent forms, especially when speaking of more powerful AI systems built using deep learning techniques that took off around 2011 and act as the foundation for more recent breakthroughs. Indeed, current deep learning–based AI systems do not yet directly alter existing catastrophic risks in any one domain to a significant degree—at least not in any obvious ways. Unanticipated present risks notwithstanding, this reality should elicit reassurance at a time of widespread anxiety about AI risks among Americans, as it gives both the government and industry an opportunity to help guide the technology's development away from the worst threats.[5] But this reality cannot encourage complacency: AI may pose very real catastrophic risks to national security in the years ahead, and some perhaps soon. The challenge for national security practitioners at this stage is to continuously monitor and anticipate emerging large-scale risks from AI as the technology rapidly evolves, often in unexpected ways, while the United States continues to ambitiously pursue AI's transformative potential. To support that effort, this report proposes four key dimensions of AI safety—the technology's novel capabilities, technical faults, integration into complex systems, and the broader conditions of its development—that will shape the risks of AI catastrophes going forward.

Terms & Concerns

ChatGPT's release launched once-obscure concerns about dangerous, high-impact AI events into the mainstream. Since ChatGPT's arrival, public discourse has seen an unprecedented focus on "existential risks"—the fear that AI could wipe out human civilization through a combination of super-human intelligence and misalignment with humanity's interests. But the groundswell of public interest in AI-related dangers has also confused the characterizations of these dangers, with experts and policymakers sometimes using terms such as "disaster," "catastrophe," and "existential threat" interchangeably, and sometimes to refer to different things.[6] The abstract nature of the threats AI poses does not help: unlike nuclear weapons, AI does not explode, and the technology's impacts—even if considerable—are often indirect. For example, if AI tools are eventually able to help develop a highly lethal pandemic supervirus for nefarious purposes, the results could prove much more devastating than any one nuclear strike, even if the crucial role of AI is more subtle.

Despite this confusion in terms and concerns, AI-related dangers have firmly established themselves in public consciousness. Fear of extreme dangers from AI motivated thousands of individuals, including many industry leaders such as Elon Musk and Apple cofounder Steve Wozniak, to issue a statement calling for a minimum six-month pause on building more advanced AI systems in the wake of ChatGPT. The statement—which suggested a government moratorium if necessary—was driven by a fear of runaway AI capabilities posing "profound risks to society and humanity."[7] Roughly two months later, a broad coalition of pioneering AI scientists and other notable figures, from OpenAI CEO Sam Altman to Microsoft cofounder Bill Gates, signed a second concise statement expressing similarly grave concerns, asserting that "mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war."[8] Numerous leading publications have published articles exploring calamitous outcomes from advanced AI systems, further escalating fears among the public and decision-makers alike about devastating outcomes from rapidly advancing AI systems.[9] A Quinnipiac poll published in May 2023 found that 54 percent of Americans now believe that AI "poses a danger to humanity," a sentiment echoed in a speech by Vice President Kamala Harris in which she declared that AI threats "could endanger the very existence of humanity."[10] Months later, the UK House of Lords published a report that identified catastrophic risks from AI as an area requiring "immediate attention" from the government, while simultaneously warning that existential risks from AI were "exaggerated and must not distract policymakers from more immediate priorities."[11]

To disentangle the large-scale threats AI poses, it is useful to distinguish terms such as "disaster," "catastrophe," and "extinction-level events." Within humanitarian and crisis response discourses, these terms have been contested for years, preventing a common set of definitions.[12] Perhaps the clearest way to organize the sometimes-overlapping concepts is by a "continuum of magnitude" in events, as proposed by Clifford E. Oliver, which distinguishes categories according to the scope of their impact and the size and complexity of the response required to manage their effects.[13]

Figure 1: Continuum of Magnitude[14]

Adapting Oliver's and others' work, a practical definition of each of these categories comes into focus.[15]

Emergencies are events at a local level that pose a risk to the life, well-being, or financial health of one or more individuals. Local officials usually have plans and processes to manage the effects of emergencies, even if their efforts are not always successful. An AI-related emergency might be a self-driving car that accidentally causes a serious road collision, requiring immediate medical attention.

Disasters are events involving multiple people, large-scale economic damage, or both. Local crisis management resources cannot sufficiently manage disasters, requiring additional support from surrounding localities, regions, or the national government. An AI-related disaster could be the malfunction of an automated oil rig system, resulting in the uncontrolled release of millions of gallons of oil into the ocean, similar in extent to the 2010 Deepwater Horizon oil spill.

Catastrophes are events of such magnitude in terms of casualties or economic destruction that they overwhelm the ability of one or more national governments' crisis management systems to fully handle their impacts, resulting in unmet critical needs, at least in initial phases. Catastrophes differ from emergencies and disasters not just in terms of size, but also in the nature of their impacts, as they elicit wide-ranging and interconnected social, political, and economic effects that cannot be managed by any one command and control system.[18] Catastrophes thus represent an overwhelming shock to existing social and governmental systems. The Federal Emergency Management Agency (FEMA) offers a similar definition for a "catastrophic incident":

> Any natural or man-made incident, including terrorism[,] that results in extraordinary levels of mass casualties, damage, or disruption severely affecting the population, infrastructure, environment, economy, national morale, and/or government functions. A catastrophic event could result in sustained national impacts over a prolonged period of time; almost immediately exceeds resources normally available to local, State, Tribal, and private sector authorities in the impacted area; and significantly interrupts governmental operations and emergency services to such an extent that national security could be threatened.[19]

An example of an AI-related catastrophe would be the use of AI to develop a deadly and highly contagious pathogen that wipes out significant swaths of a national population, similar to the effects of the 1918 Spanish Flu.

Extinction-level events are cataclysmic in scope, threatening to wipe out the human species, such as if a large asteroid were to collide with Earth in such a way that the world became uninhabitable. The risks of such events occurring are often referred to as **existential risks** in public discourse, but definitions of existential risks also include a broader set of scenarios in which an event may not totally wipe out human life, but would "permanently and drastically curtail" the potential of intelligent life or lead to an irreversible collapse of civilization.[21] Definitionally, coping with the impacts of such events is beyond the capacity of humanity's collective mechanisms for crisis management. In recent years, existential risks have garnered considerable attention from academics, in part inspired by the work of Oxford philosopher Nick Bostrom, who has argued for them to receive far greater attention and resources.[22]

Some confusion in the terminology around AI risks stems from a sizeable focus on advanced future AI systems as an existential threat to humanity. Several theorists, including Bostrom, have posited that if one or more AI systems could surpass human intelligence, a failure to fully align the systems' interests with human flourishing could threaten civilization.[23] Some proposed scenarios suggest that these risks could play out very quickly as a singular extinction-level event, while others suggest a more gradual process of extinction or crippling disempowerment as humans cede agency and economic vitality to a superintelligent AI system.[24]

Though existential risks have been a prominent issue in public discourse around AI since ChatGPT's release, the characterization of these risks remains hotly contested by experts. Some view addressing existential risk from AI as a pressing priority due to the very rapid progress of AI in recent years, the unsolved challenges of reliably controlling AI behavior, and the observed ability of AI systems to produce considerable effects in societies already.[25] Other experts dismiss the likelihood or intrinsic dangers of developing AI with superhuman intelligence and express concern that the focus on existential AI risks is distortionary or even dangerous in itself. In this view, the specter of existential AI risks is fueled by commercial incentives to hype AI products and sideline more immediate social and legal issues associated with the technology, in combination with Luddite and apocalyptic impulses in society that have tended to accompany periods of technological acceleration and social change.[26] So great are the disagreements between technologists on the issue that their cantankerous debate was compared to a religious schism in *The Economist*.[27] In any case, because so many other publications have focused on existential risks from AI, this report will focus instead on catastrophic risks and their relationship to national security.

Emergency: *After the 2016 fatal crash of this Tesla Model S near Williston, Florida, investigators at the National Transportation Safety Board found that overreliance on the vehicle's autopilot mode, flaws in the autopilot system design, and user error all contributed to the crash.*[16] *(National Transportation Safety Board via Flickr)*

Disaster: *The Deepwater Horizon Oil spill released four million barrels of oil over 87 days, causing billions in damages.*[17] *(U.S. Coast Guard via Flickr)*

Catastrophe: *If scaled to today's population, the Spanish Flu would have killed approximately 70–150 million individuals globally, or roughly 2–6 times as many as the COVID-19 pandemic.*[20] *(GPA Photo Archive/National Archives via Flickr)*

Extinction-level event: *A large asteroid colliding with Earth could cause devastation on initial impact through kinetic energy but may also throw massive amounts of soot and dust into the atmosphere that would block out sunlight for long enough to kill off plant life and collapse food chains. (Marc Ward/ Stocktrek Images via Getty Images)*

Clarifying Catastrophe

Although this continuum of risks can help clarify some of the confusion among terms, it comes with important caveats: as the continuum implies, the boundaries between categories are inexact. Events can straddle categories, such as a relatively large disaster or a comparatively limited catastrophe. Crises of various kinds can, and often do, cross borders and classifications as they evolve. Likewise, crisis mismanagement can turn lower-magnitude events into larger and more dangerous ones, for instance, if a poor healthcare system response allows a local outbreak to grow into an epidemic or pandemic. Contributing to the confusion between catastrophic risks and existential risks is the fact that a catastrophe could in principle snowball into an existential threat—whether of its own accord or through mismanagement—even though the gulf between even a very large catastrophe and a true extinction-level event is much greater than many suppose.[28]

By focusing on acute public safety risks from discrete events, such as pandemics or industrial accidents, this continuum does not capture a wide range of more diffuse AI harms in society. For example, risks related to labor automation and job displacement, systemic bias and fairness, mass surveillance, widespread disinformation, and other diffuse harms can—and in some cases already do—affect millions of people and have widespread economic, political, and social impacts. Because these harms occur as numerous incidents that are part of a larger, ongoing trend rather than a large, discrete event, they typically are not characterized as "disasters" or "catastrophic events." This does not make them any less important. In fact, the insidious nature of their harm can sometimes make them more challenging to address relative to discrete large-scale events that have a clear public safety impact. These types of diffuse harms require attention and intervention, both for harms that exist today and for those that may materialize in the future, such as rapid-onset labor automation. The types of interventions needed for such issues, however, usually differ considerably from those required to address catastrophic risks, in part because they tend to be more politically fraught and thus necessitate greater social deliberation and coalition building. Given these differences, and the fact that these issues have been covered extensively elsewhere, they are not addressed in this report.[29]

Because the definitions of emergencies, disasters, and catastrophes are tied to government responses, there is variability between countries as to which events fall into which categories. For example, a hurricane that hits a large area may be classified as a disaster that can be dealt with by regional authorities in the United States, but amount to a catastrophe in a less-developed or smaller nation less equipped to cope with the hurricane's effects, such as the devastation wrought by Hurricane Mitch in Honduras in 1998.

Relatedly, AI's impact on "epistemic security"—societies' ability to effectively process and act on information—can also impact the degree to which a state can manage crises. For example, critics on both the right and left have pointed out how eroded epistemic security adversely affected the government's COVID-19 response.[34] Experts fear that current and future AI bots and algorithm-driven information echo chambers could so degrade states' crisis response systems that they indirectly contribute to catastrophes not by exacerbating the inciting event itself, but by frustrating the state's ability to respond effectively.[35]

In some instances, states themselves can be the source of catastrophes, even if the definition of "catastrophe" is tied to states' ability to respond. For example, China's Great Leap Forward was arguably the deadliest catastrophe of the past century, as Mao

War and Catastrophe

Wars share many of the attributes of catastrophes but are conventionally treated as a distinct—if often interrelated—issue for a few reasons. For one, whereas catastrophes are primarily managed by crisis management systems, wars are primarily managed by states' defense organs, even if there is often overlap in both cases. Additionally, managing war tends to involve a set of strategies and techniques that aim to achieve specific political ends with their own well-developed corpus of thought, independent of crisis management.[30] The techniques for managing catastrophes, by contrast, tend to be directed more toward narrower goals of curbing casualties and economic losses, and returning to a state of relative normalcy. For this reason, and due to the fact that AI disruptions to war have been handled extensively elsewhere, this report does not focus on war itself as a genre of catastrophe.[31] Nonetheless, catastrophes can initiate or contribute to the outbreak of wars, as some have argued regarding the drought that preceded the Syrian Civil War.[32] Likewise, some acts of war, especially those that involve civilians and thereby initiate crisis management systems, would also qualify as catastrophes, such as the destruction of a city by air raids, a nuclear strike on a nation's homeland, a war-induced famine, or cyberattacks on critical infrastructure. The terrorist attacks of September 11, 2001, for example, might simultaneously qualify as a catastrophe and an act of war.

In October 1998, Hurricane Mitch wrought devastation across Latin America. In Honduras, about one-third of the population was affected in an event that Honduran President Carlos Flores Facusse estimated set the country back 50 years in its development.[33] (Robert Ford via Getty Images)

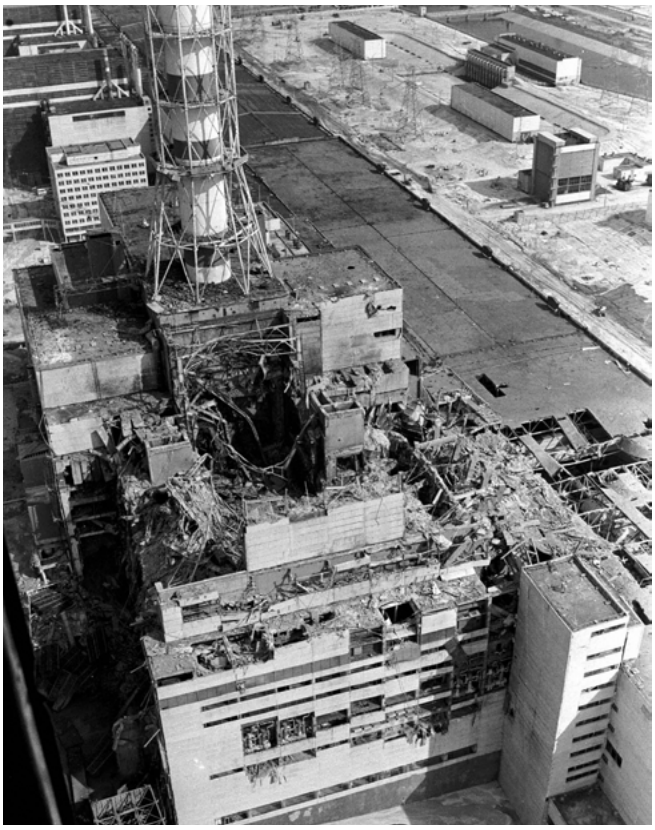
Zedong's state-led drive to transform China from an agricultural society into an industrial powerhouse inadvertently caused between 23 and 55 million deaths from starvation.[36]

Finally, what catastrophic effects and successful management of those effects look like is, to a degree, subjective. FEMA's definition of a catastrophic incident, for example, includes extraordinary disruptions to national morale as a sufficient feature of an event to qualify as a catastrophe. While the definition proposed in this report is more narrowly concerned with the magnitude of casualties and/or economic destruction as core indicators of a catastrophe, it is worth considering that alternative definitions may more highly prioritize other features that often go hand in hand with large-scale losses of lives or economic vitality.

The Priority of Addressing AI Catastrophes

The risks of AI-related emergencies, disasters, catastrophes, and extinction-level events are all worthy of attention as AI technologies mature, but catastrophic risks are particularly relevant to national security policymakers for several reasons. For one, given that the effects of catastrophes overwhelm the response capacities of all sub-national authorities, national security practitioners bear much of the primary responsibility of addressing such events. Moreover, one feature of catastrophes that distinguishes them from both emergencies and disasters is the lack of a state's ability to fully manage the effects of the event. This means that beyond causing a devastating loss of life, economic vitality, or both, catastrophes can also threaten the long-term health, security, and stability of the state itself. History is rife with instances of states' decline or collapse in the wake of catastrophe, from Athens's plague-induced deterioration during the Peloponnesian War to the collapse of Minoan society in the wake of a devastating volcanic eruption.[37] More recently, the Managua earthquake of 1972, the Bhola cyclone of 1970, and the Ethiopian drought of 1973–74 all represent instances of natural catastrophes contributing to regime change.[38] Though such catastrophic events are rare, they do occur—despite individuals' well-attested tendency to underestimate their likelihood and impacts, a phenomenon known as "normalcy bias."[39] Given catastrophes' often-dire consequences for states and societies, reducing their likelihood and planning for their effects is of utmost importance to policymakers, particularly when technological acceleration introduces new risks.

In scientifically advanced societies, powerful technologies can often catalyze catastrophes, as in the case of the nuclear meltdown at Chernobyl. The plant meltdown led to the uncontrolled release of 400 times as much radioactive fallout as the U.S. nuclear bomb dropped on Hiroshima, and famously contributed to the political collapse of the Soviet Union, as Mikhail Gorbachev himself acknowledged.[40] In this sense, technological advancement can act as a double-edged sword for developed societies. While naturally occurring events—in the form of plagues, famines, earthquakes, volcanic eruptions, tsunamis, or hurricanes—have historically been the primary source of large-scale catastrophes for most societies, technological advances have helped blunt many of the worst impacts of natural events. At the same time, however, growing technological capabilities have dramatically increased the risks and scope of man-made catastrophes. This trend is evident in war, where technological advancement has enabled the creation of ever-more destructive weapons—from sharp stones to crossbows to machine guns and finally nuclear



A helicopter view of the destruction at the nuclear plant in Chernobyl, Ukraine, just a few days after the meltdown in April 1986. (Vladimir Repik/AFP via Getty Images)

weapons with the capacity to kill billions. Outside of advances in weaponry, too, the trend holds. Humanity's growing ability to wield increasingly powerful technologies creates the potential for ever-greater catastrophes from inadvertent civilian applications of technology—from possible leakages of dangerous pathogens to nuclear reactor meltdowns to human-caused ecological collapse.[41]

If war is any indicator, the overall impact of technological advancement ostensibly heightens the relative priority of mitigating catastrophic risks relative to disasters and emergencies. Whereas the exponential progress in weapons' destructive capacity has been partially offset by advances in medicine and defense technologies in terms of fatalities in conventional conflicts, states still have limited options to manage the risks of more extreme events, such as nuclear strikes.[44] A similar dynamic could be at play in civilian uses of powerful technologies. Technology has provided tools to more effectively manage emergencies and disasters of a more limited nature. But technology has simultaneously escalated the dangers of large-scale catastrophes by unleashing extremely destructive forces upon society with effects of extraordinary magnitude, such as nuclear reactor meltdowns and sophisticated bioweapons. This dynamic is most acute when technologies are in their infancy—before the risks are fully understood and corrective measures are established over time through trial and error.[45]

With the partial exception of AI-powered autonomous weapons, the destructive potential of AI may not be as readily apparent as that of some other technologies. Even if AI tools do not explode like nuclear bombs, AI systems' more subtle and complex hazards may be no less profound. Like electricity, AI is a general-purpose technology—able to be used in a vast array of applications—and is being rapidly integrated into complex, delicate systems from healthcare to global logistics, as well as unlocking scientific breakthroughs in multiple fields. Since 2019, private investment in AI development has exceeded $100 billion per year and may well rise further in the near term, accelerating AI progress and deployment.[46] The increasing capabilities of AI systems, whose inner workings are often inscrutable to human oversight and sometimes superior to human abilities, means that sophisticated AI tools in combination with other systems and technologies could significantly alter the risk profile of hazardous domains. The speed of AI deployment, the diversity of potential applications, and the quickly growing capabilities of AI models all lend themselves to heightened catastrophic risks in a variety of fields.

Some incidents have already demonstrated a proof of concept for possible catastrophic risks in which AI plays a role. AI tools have demonstrated the ability to aid in the design and manufacture of chemical weapons, suggesting a potential future in which nonstate actors can more easily develop and launch chemical—and perhaps eventually biological—attacks.[47] In 2010, algorithmic trading laid the foundation for a "flash crash," causing a trillion dollars to be temporarily wiped out of the stock market.[48] With the chair of the Securities and Exchange Commission (SEC) warning that AI "will be the center of . . . future financial crises," far more debilitating crashes may well be on the way.[49] Automated military systems used in nuclear command and control have also suffered failures and false alarms—incidents that some fear could portend AI-induced nuclear catastrophe scenarios in the near future.[50] As government leaders grapple with the dangers AI poses, it is important to better understand potential risks of AI catastrophes.[51]

Given these precedents, corollary fears of AI-enabled bioterrorism, runaway cyberattacks, financial meltdowns, and nuclear misfires naturally represent the

Figure 2: Growth in Weapon Lethality over Time

As technology has improved, the destructive capacities of weapons have increased over time. Though theoretical, Trevor Dupuy's attempt to quantify the lethality of weapons based on range, rate of fire, accuracy, reliability, radius of damage, and other factors gives some indication of the growth of destructive capacity in weapons driven by technological advancement.[42] Alexander Kott's efforts to explore the performance power of direct-fire weapon systems over centuries suggests a similar story of the exponential growth of destructive power.[43]

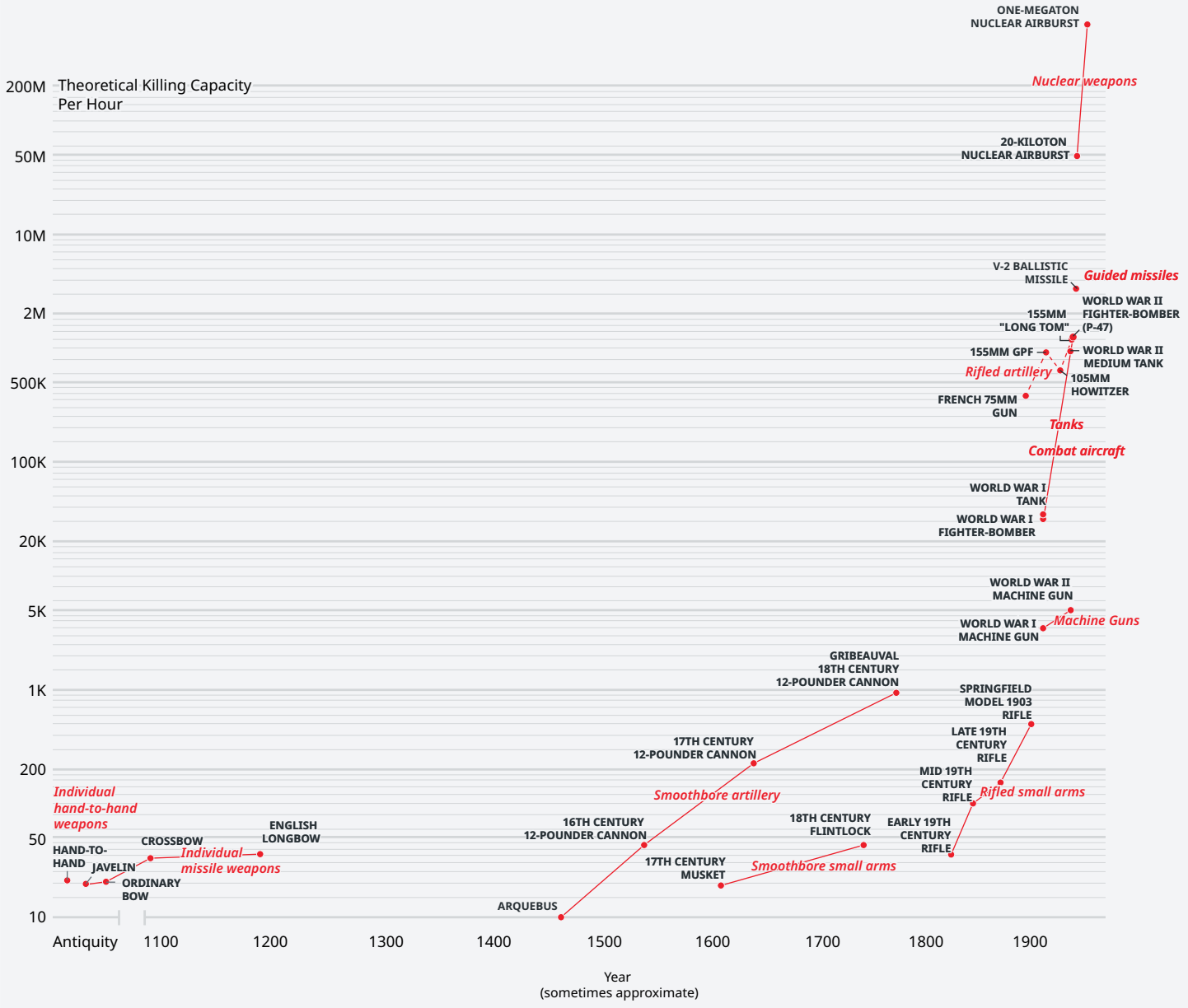


Chart by Trevor N. Dupuy, *The Evolution of Weapons And Warfare*. Adapted by Bill Drexel and Caleb Withers; Design: Melody Cook/CNAS.

scenarios that garner the most attention for near- to medium-term catastrophic risks. There are very good reasons to focus attention on each case. But the clear recognition of these specific fears may also help to curb their likelihood. By contrast, responding to more unexpected developments that attract less attention over time may ultimately prove more challenging, highlighting the importance of building awareness of and resilience to a more holistic set of AI safety dynamics.

Indeed, these and other risks of catastrophe in high-impact domains, such as biosecurity, cybersecurity, finance, autonomous weapons, high-risk industries, critical infrastructure, and nuclear command and control are far from static regardless of AI developments, and depend on scientific, technological, political, and social changes in each domain. Assessing AI-related risks in any one domain thus involves the interplay between two moving targets: the changing risks of the domain itself and rapidly developing AI capabilities. Additionally, the relationship between scientific progress in AI and high-risk domains such as biotechnology or cybersecurity often exhibits a synergistic effect, as new capabilities in one field can unlock new capabilities in the other. This amplification effect is referred to as "technological convergence" and adds another layer of complexity to characterizing risk in these domains.[52] Due to this complexity, unforeseen developments in any domain could alter pathways to catastrophe in unpredictable ways. Despite this uncertainty, the considerable work that has been done on how technological progress interacts with safety risks, and the sub-discipline of AI safety in particular, can help illuminate how AI development can impact catastrophic risks with national security implications.

A final reason why AI catastrophes are worthy of attention is that in addition to potentially exacerbating the chances and severity of catastrophes in a variety of domains, AI could also make it even more difficult for states to manage their effects. As mentioned previously, some experts fear that AI tools such as deepfakes, LLMs, and more sophisticated recommendation algorithms could considerably degrade societies' information environments, in turn degrading their crisis response capabilities.[53] In this view, a combination of more convincing, personalized, and abundant mis- and disinformation created from AI tools and greater media polarization from siloed, AI-fueled media subcultures could make citizens more susceptible to false narratives. Such an environment would inhibit the ability of states to make and execute decisions in times of crisis, and would erode public trust generally over time. Already, LLMs have shown potential in lowering the cost and enhancing the quality and scale of disinformation operations, and deepfakes are being deployed in high-profile cases to influence consequential political processes.[54] While much of the work on these issues has focused on risks to open, democratic media ecosystems, AI tools could have parallel effects in autocratic systems, albeit through different means. Rather than sowing distrust and confusion, autocrats' use of AI to bolster propaganda and censorship could exacerbate the challenges of information distortion that plague autocratic regimes, in which critical information fails to reach autocratic leaders, who in turn make poor decisions that can exacerbate or initiate crises as they begin to believe their own propaganda.[55] The Great Leap Forward—the largest catastrophe of the past century by number of casualties—was in large part fueled by such information distortion, suggesting that despite the outsized focus on open societies, AI's impact on the information ecosystems of closed societies may be more severe in terms of catastrophes.[56]

The increasing capabilities of AI systems, whose inner workings are often inscrutable to human oversight and sometimes superior to human abilities, means that sophisticated AI tools in combination with other systems and technologies could significantly alter the risk profile of hazardous domains.

AI's impact on information environments is a risk factor that differs in kind from the AI safety dynamics that are the primary subject of this report insofar as it acts as an overarching concern that could affect AI-related crises in any domain, as well as crises that emerge independently of AI tools. Given that a state's relative ability to respond to a disruptive event ultimately determines the extent of the event's impacts—and that catastrophes are often the result of state mismanagement of smaller-scale disasters—the influence of AI-powered media degradation on catastrophic risks could be considerable.

Catastrophic Risks and Dimensions of AI Safety

With the exception of a handful of specific proposed scenarios discussed in the following sections, for the most part AI catastrophic risks of relevance to national security are still taking shape. Because advanced AI applications in high-impact domains are mostly in their infancy, much of the concern about AI catastrophes today is prospective—a well-informed intuition that the vast power of AI likely *could* result in tremendous hazards once applied to consequential arenas, even if the largest risks have yet to materialize. Nonetheless, given the rapid pace of AI advancement and considerable scope for the impacts of AI in national security, considering how AI safety could impact catastrophic risks as the technology develops is indispensable—offering the opportunity to guide the technology's development toward safety and stability to the extent possible, rather than retroactively addressing severe risks after they have emerged. A clearer awareness of the underlying dynamics driving catastrophic risks related to AI can help build resilience and reduce the chances of experiencing a major AI catastrophe.

In the service of helping to shape preparation for catastrophic risks of AI even as the technology develops, this report proposes four broad dimensions of AI safety as they relate to catastrophic risk.

These dimensions distill the insights of a range of both AI-specific and broader literature on safety and risk, aiming to be flexible enough to apply across a wide range of domains.[57] To further explore these categories, subissues are identified in each area. Though presented independently, in practice these issues often overlap. After clarifying each subissue, its relevance to catastrophic risks in national security is examined. Although existing incidents and precedents are cited where possible in these explorations, many of the scenarios proposed are largely hypothetical, and some may not be relevant for many years to come, if ever. Additionally, these themes are narrowly focused on understanding the set of issues that contribute to AI's catastrophic risks, and do not include solutions. In all cases, researchers and engineers are working to address these dynamics, but to recount that work is beyond the scope of this report. The issues considered here aim not to be exhaustive, but to provide a foundation with key examples and references to broader safety literature as a means to more holistically assess how AI can shape catastrophic risks as the technology evolves and is increasingly built into consequential systems.

Dimension	Question
New capabilities	What dangers arise from new AI-enabled capabilities across different domains?
Technical safety challenges	In what ways can technical failures in AI-enabled systems escalate risks?
Integrating AI into complex systems	How can the integration of AI into high-risk systems disrupt or derail their operations?
Conditions of AI development	How do the conditions under which AI tools are developed influence their safety?

It is important to note that this exploration is not intended to provide a full risk management assessment for any particular scenario, which is traditionally a three-step process:

1. Assessing risk as a factor of likelihood (including both threats and vulnerability) and consequence.
2. Considering mitigations for threats, vulnerabilities, and consequences.
3. Prioritizing mitigations that most reduce overall risk.

Given how broad and fast-moving AI applications are, and the fact that the rollout of advanced AI capabilities across domains is largely still in its infancy, accurately assessing the full range of the most likely AI catastrophic threats, vulnerabilities, and consequences is simply not possible. Threats and vulnerabilities will vary widely between domains, and will evolve over time depending considerably on how deeply AI tools are integrated into high-impact systems that have the potential to disrupt society.[58] Systems associated with biological security, cybersecurity, financial security, militaries, high-risk industries, and critical infrastructure are the most obvious candidates, but there may well be others. Given the immense promise that AI systems hold, there is good reason to believe that they may eventually become highly integrated into any or all

of these domains. But the timing and conditions under which such integration occurs is an open question and will vary.

In most cases, trying to assign specific likelihoods to not-yet-developed systems would be premature. What may prospectively seem like the most obvious high-risk scenarios in a sector are often also the most likely to be addressed early, meaning that the very act of clearly identifying a specific pathway to catastrophe may reduce its likelihood of occurring. But even predicting "likely" scenarios early is a challenge: reality so often proves stranger than fiction, contingent on unpredictable forces and extraordinary courses of events.[59]

Rather than providing risk management assessments themselves, this report aims to help lay a foundation for future risk management assessments, which will require continuous updating and more granular attention to specific scenarios based on a range of variables, including:

- Risk types: misuse (e.g., AI-enhanced bioweapons), accidents (laboratory leaks), or structural issues (widespread poor biosafety controls due to insufficient safety research)[60]
- Specific domains (cybersecurity, biosecurity, finance, nuclear stability, autonomous weapons, high-risk industries)
- Actors (lone wolves, terrorist organizations, states, corporations)
- Incentives (terror, prestige, profits, regulatory environments)
- Types of AI models ("narrow" models vs. general-use or "foundation" models)

To examine more specifically how these AI safety dimensions manifest in a particular domain of interest, this report will be paired with a follow-on report, *AI and the Evolution of Biological National Security Risks*.

A final word of caution is in order before delving into the many dynamics that could contribute to AI catastrophes. Restricting this primer only to the possible dangers stemming from AI runs the risk of fostering an excessive fixation on what could go wrong, rather than an affirmative vision of what could go right. Readers should avoid this distortion. The opportunity costs of failing to proactively pursue AI development, while impossible to measure, could be severe. As societies become more complex, leveraging AI to help manage their complexity will likely be an overall boon to reducing catastrophic risks—not to mention the tremendous potential of AI to enhance America's economy and national security. Relatedly, falling behind China, an adversary with the stated goal of supplanting the United States' leading position in AI, also represents a severe risk.[61] As further described below, not only would Chinese preeminence in AI grant Beijing strategic economic and military advantages over the United States and help bolster autocratic rule around the world, it would also greatly exacerbate the likelihood of AI catastrophes generally.[62] For these reasons, it is imperative that the United States continue to boldly pioneer the development of AI technologies. Highlighting the dynamics of AI catastrophic risks is in the service of that goal—*not* an admonition against ambitiously building powerful, effective AI tools.

New Capabilities

New capabilities from AI tools can have dangerous impacts across a range of domains, either directly from AI systems themselves or from AI-related breakthroughs in adjacent scientific or technological domains. These dangers are most prominent in relation to cyber, epistemic, biological, and chemical security, where sudden new capabilities could have dramatic effects, and in some cases disrupt existing deterrents and technical or financial barriers that serve to mitigate the risks of catastrophe.

DANGEROUS CAPABILITIES

A range of AI tools exhibit hazardous capabilities of relevance to several high-risk domains that experts anticipate will become only more powerful as the technology progresses. Models can produce mis- and disinformation at scale and with increasing quality, posing a threat to societies' information ecosystems.[63] In biological and chemical applications, AI systems have shown potential in helping to develop weaponizable chemicals or pathogens (although not necessarily aiding actors any more than existing tools, and in different ways depending on the preexisting expertise of users).[64] Cybersecurity professionals also see growing use of generative AI in phishing attacks and anticipate more sophisticated AI capabilities on the horizon.[65] Additionally, seemingly benign AI-enabled capabilities could have hazardous implications, such as advancements in material science, jet propulsion, or other fields that could be repurposed for weapons use.

Implications for Catastrophic Risk

Experts have warned that dangerous capabilities from emerging AI tools could raise the likelihood, severity, or both, of catastrophic attacks in both biosecurity and cybersecurity. In the former, general-use foundation models could lower the barriers to entry for bad actors seeking to build or procure high-impact bioweapons, while AI-powered "biological design tools" may eventually help craft more strategic or deadly biological agents.[66] Cybersecurity experts have likewise warned that AI tools could make cyberattack capabilities more broadly accessible, and enhance the quality and sophistication of advanced cyberattacks, potentially targeting critical infrastructure with catastrophic effects.[67] Cyberattacks could also target emergency response communications systems, further exacerbating the impacts of crisis events.

But sudden, new capabilities related to AI advancements could also exacerbate the risks of catastrophe in less direct ways. As previously mentioned, the use of LLMs and other tools may degrade a state's ability to cope with disasters or catastrophes by facilitating higher-volume and better-quality misinformation and disinformation at scale.[68] Additionally, AI technologies' tendency to usher in sudden breakthroughs in a wide variety of scientific subfields could lead to the sudden introduction of strategically disruptive new capabilities that escalate the chances of miscalculations in high-stakes domains, including conventional or even nuclear deterrence.[69] In such cases, where stability is predicated on a degree of confidence about capabilities on both sides, sudden new capabilities can upend the delicate equilibrium of actions and reactions that is fundamental to stability. For example, nuclear stability could be greatly impacted if one country unexpectedly developed an AI tool able to crack encryption systems protecting nuclear command and control systems abroad, or if AI-enabled breakthroughs in nuclear delivery systems offered one country a sudden, significant advantage over adversaries. Although these are provocative examples, more subtle gradations of this dynamic are possible across a range of domains.

EMERGENT CAPABILITIES

The capabilities of foundation AI models have steadily increased over time alongside the exponential growth of compute used to train them.[70] However, specific capabilities can emerge suddenly, improving sharply from minimal to strong competence as models are scaled. These capabilities can emerge at seemingly unpredictable points and without specific encouragement from model developers—although researchers have contested the degree to which such capability jumps are truly surprising, or simply a mirage originating from the methods used to measure capabilities.[71] In practice, this means that the specific capabilities of newly developed models often cannot be fully anticipated before training: disruptive or destructive capabilities may fall into the hands of developers who were not seeking or preparing for them, posing challenges for the management of strategic and potentially risky applications.[72]

Implications for Catastrophic Risk

Uncertainty about the timing and nature of dangerous AI capabilities as they emerge—often as unintended byproducts of the pursuit of other capabilities—further complicates states' abilities to mitigate catastrophic risks from emerging AI systems. The warning signs of emerging risks that accompany more incremental, predictable technological development may be less regular or less pronounced in the case of AI, making it difficult to develop safeguards ahead of systems' proliferation.

LATENT CAPABILITIES

AI models' capabilities may not always be detected by their creators, such that dangerous capabilities may only become known when stumbled upon by others—including, perhaps, malicious actors. For example, foundation models, including large language models, are primarily trained on relatively simple tasks, such as predicting the next word (or part of a word) in a sequence of text. But these simple objectives have given rise to a vast array of practical capabilities—more than can be exhaustively tested for.[73] Researchers continue to discover new methods to elicit significant performance improvements that the models' creators did not initially anticipate, with minimal additional training, through methods such as fine-tuning on tailored datasets, knowledge distillation from larger models, or prompting techniques such as chain-of-thought reasoning.[74] Language models have even shown the ability to learn representations that extend beyond language tasks, proving useful for domains such as image classification and protein fold prediction.[75] Even narrow models regularly demonstrate "transfer learning," where knowledge gained from one task proves useful in others with varying degrees of relation.[76]

Implications for Catastrophic Risk

As AI systems proliferate, undetected latent capabilities could contribute to bad actors' ability to initiate

catastrophic events that the models' developers may not have imagined possible. For example, researchers at the North Carolina–based company Collaboration Pharmaceuticals inverted one AI tool designed for discovering therapeutic molecules as a thought experiment for a security conference—and were surprised to find that within six hours, their inverted tool had proposed 40,000 candidate chemical compounds that might be viable chemical weapons, including several known agents that were not included in the model's training data.[77] Though the adjustment was easily made, the researchers did not anticipate that their model could be so readily misused to such great effect. Similar incidents in other domains, especially where the AI models in question are publicly released, could expand the capabilities of bad actors.

Technical Safety Challenges

Technical safety challenges intrinsic to AI will continue to create vulnerabilities as AI tools increasingly integrate with sensitive systems. Though often arcane, technical faults in AI systems can have dire consequences: for example, errors in image recognition systems in self-driving cars have already led to several fatalities. AI engineers are currently working to address these issues, though the degree to which they will ever be fully "solved" is an open question. As in many technical systems, there will likely be incremental improvements to these issues that can always reemerge as AI systems develop more powerful capabilities and are applied in new contexts.

ALIGNMENT AND SPECIFICATION GAMING

For AI systems tasked with achieving particular goals, specifying objectives that accurately reflect their designers' intentions remains a persistent challenge.[78] Such systems have been known to find various ways of "hacking" the specified goals, often by violating unspecified or underspecified rules that might seem obvious or common sense to their programmers and are not explicitly encoded into the system's instructions. This is known as specification gaming. For example, one AI system instructed to win a boat race video game discovered that it could maximize its points by driving in chaotic circles through reward tokens rather than by completing the race.[79] The effect is similar to how someone might exploit the letter of the law rather than following its spirit. Although the boat example is innocuous, as AI systems integrate with more complex systems, the consequences of specification gaming can become much more severe. Taken to the extreme, some fear that future, superintelligent AI systems misaligned with human interests could pose catastrophic or even existential risks. These risks are highly speculative, and expert opinions range widely

M55 rockets containing the VX nerve agent are examined prior to their destruction in accordance with the Chemical Weapons Convention. VX was among the 40,000 candidate compounds proposed by Collaboration Pharmaceutical's inverted AI system. (Program Executive Office, Assembled Chemical Weapons Alternatives via Flickr)

about when or if so-called "artificial general intelligence" (AGI) or "superintelligence" could emerge, but the concern has gained traction among many in leading labs and some high-level political leaders.

LOSS OF CONTROL

Operators could lose control of AI systems for a variety of reasons, posing the risk of a "runaway" or rogue system causing damage in high-impact systems. Though this issue is often associated with aforementioned AGI or superintelligence concerns, it need not be: costly loss of control could occur in comparatively simple systems.[80] Consider, for example, how the 2017 NotPetya cyberattack spread uncontrollably around the world at the cost of more than $10 billion, attacking systems in hospitals, global shipping companies, and factories. It even rebounded on the originating country, Russia, by hitting Rosneft, a state oil company.[81] Emerging AI capabilities hold the potential for still more sophisticated autonomy, which could mean more dynamic—and dangerous—risks from loss of control.

ROBUSTNESS

AI models are deemed "robust" when they consistently perform well across a wide range of conditions, especially those that deviate from their training data sets.[82] Achieving robustness can be challenging. Strategies to enhance robustness can include diversification of training data; techniques to reduce overfitting to training data, such as ensemble systems that use multiple models in parallel to improve accuracy in their determinations; and provision for fallbacks (such as seeking human input) when encountering anomalous situations. In some cases, performance may simply degrade in new contexts. But in others, AI systems may retain their capabilities while "misgeneralizing" their goals—or employing coherent strategies in pursuit of incorrect objectives.[83]

CALIBRATION

The calibration of AI systems reflects how well the confidence in their determinations corresponds to correctness. Calibration can help ensure that AI systems know when they can act confidently, and when to seek assistance or avoid high-stakes decision-making.[84]

The calibration performance of an AI model can be quantified by comparing its predictions with outcomes. Measuring calibration can be more complex than it might initially appear, however, and high performance on one method of gauging calibration does not guarantee high performance on another.[85] Calibration and robustness often go hand in hand, as calibration can be a particular problem in situations outside of training distributions. Conversely, well-calibrated models can help identify and mitigate the risks associated with poor performance in scenarios that deviate from training distributions.

Implications for Catastrophic Risk

Issues such as alignment, specification gaming, loss of control, robustness, and calibration are all integral to ensuring that AI systems behave reliably and according to their intended purposes. To the degree that AI systems are used to help manage high-stakes processes, insufficient attention to any one of these issues could contribute to catastrophic outcomes.

A variety of military contexts could be applicable to these issues, most obviously if powerful lethal autonomous weapons misfire under politically fraught conditions. Such a malfunction could catalyze political or military escalation with potentially catastrophic consequences, though such a course of events would ultimately be determined by subsequent policy and strategy decisions. The U.S. military has been proactive in promoting rigorous standards for AI across its operations to avoid such scenarios, and it does not currently field lethal autonomous weapons systems that would initiate such a chain of events. However, as lethal autonomous weapons become more sophisticated, the likelihood of consequential accidents or inadvertent escalation from technical AI challenges rises, particularly if rigorous standards are not adopted more widely. Weapons aside, AI systems entrusted to help manage highly complex military logistics and maintenance systems could also have consequential impacts in the case of technical failures, in some circumstances potentially contributing to the chances of a catastrophe.

Beyond militaries, AI systems used to help manage high-risk systems in nuclear energy, chemical plants, biosafety level 4 (BSL-4) labs, cybersecurity, transportation systems, or elsewhere could also go awry in catastrophic ways due to technical flaws, and require appropriate mitigation measures.

ADVERSARIAL ATTACKS

Adversarial attacks can induce AI systems to err due to deliberately crafted malicious inputs. These inputs are often designed to be imperceptible to humans, but with subtle changes that specifically target the AI system—for example, adding a few pixels to an image of a cat to make it register as a dog.[86] Adversarial manipulation can also be used to extract sensitive information from an AI system or its training data. Foundation models face additional challenges in withstanding adversarial threats. LLMs,

for example, can be coaxed in plain English to produce outputs that contradict their safety training.

Attacks can be even more powerful if the aggressors have influence over a model's training. For instance, an attacker might insert "poisoned" data into a training set to make a model behave differently in certain situations—either through actual infiltration, or through uploading information to the internet that might then be scraped by AI labs. There are currently no foolproof defenses against adversarial threats, especially without impacting the performance of AI models. Specific attack methods and defenses continue to develop in a cat-and-mouse game.[87]

Implications for Catastrophic Risk

On a limited scale, researchers have already demonstrated how adversarial attacks can have dangerous effects in the real world. Two of the most notable examples include inducing an autonomous car to swerve into an oncoming-traffic lane through carefully applied small markings on a road's surface, and using specialized glasses to spoof facial recognition security cameras and evade recognition or allow impersonation.[88] Though these forms of adversarial attacks are unique to AI systems, many of their associated risks parallel those of cybersecurity vulnerabilities: whether hacking conventional computer systems or hacking AI tools, both methods could in principle allow adversaries to induce malfunction in strategic systems such as critical infrastructure. To the degree that high-impact systems such as critical infrastructure begin to use AI to help manage their complexity, so too will such systems be vulnerable to adversarial attacks.

EXPLAINABILITY AND INTERPRETABILITY

Advanced AI systems are increasingly built using deep learning models, which include many-layered "neural" networks with inner workings that can be very difficult to explain or interpret.[89] As deep learning models become larger and their performance improves, the difficulty of understanding their inner workings becomes greater. A direct trade-off often occurs between performance and explainability, as models offer increasingly strong performance without developers or users fully grasping how or why they make the decisions or outputs they do.[90] Over time, some worry that continued reliance on highly useful—but insufficiently understood—machines will lead to precarious accumulations of "intellectual debt" that can easily go awry as the difficulty of anticipating and understanding unexpected behavior compounds.[91]

Implications for Catastrophic Risk

Technical challenges in explainability and interpretability are unlikely to directly lead to a catastrophic event but have indirect relevance worthy of note. In instances of accidental technical malfunctions that might lend themselves to dangerous escalation, an inability to demonstrate how and why a system malfunctioned could exacerbate mistrust and accelerate retaliatory action. Think, for example, of an AI-powered missile defense system erroneously firing on an adversarial nation.

Conversely, if AI tools remain largely inexplicable, their integration into an ever-wider set of national security–related systems would represent yet another arena for destabilizing gray zone operations, offering ample room for denial to cover subtle strategic attacks. For example, an adversarial actor could shut off access to crucial energy management systems by exploiting unknown vulnerabilities in an AI system helping to manage electricity grids. The adversary would have much greater leeway to claim that the AI system simply malfunctioned if the nature of its malfunction remained opaque. Likewise, if an AI system were known to be opaque, an adversary could plausibly claim to be the source of a malfunction or to have the capability to cause a malfunction even if not true.

Emerging AI capabilities hold the potential for still more sophisticated autonomy, which could mean more dynamic—and dangerous—risks from loss of control.

Finally, AI systems' lack of explainability makes it far more difficult to troubleshoot and address other technical issues reliably, posing long-term challenges to ensuring AI systems' integrity, including in high-risk domains.

Integrating AI into Complex Systems

Integrating AI into complex systems presents an added layer of safety challenges that could have catastrophic effects. This aspect is often overlooked due to greater attention on the risks of new capabilities and technical issues, but there is good reason to believe that how AI tools are integrated into broader systems—including how human operators respond to systems in practice—will be a key part of the risk profile for AI in high-risk

domains.[92] Historically, mundane lapses of judgment and human operator errors have often been at the root of many automation-related tragedies, even if new technical capabilities and related technical flaws tend to dominate safety discussions. In light of this reality, "human-machine teaming" has emerged as a field of study to try to discern how the design of automation-enabled systems can best work with the cognitive and emotional particulars of diverse operators. Moreover, even if there are no flaws with how users or operators engage with AI tools, the introduction of powerful new automation apparatuses into broader, complex ecosystems can, and often does, produce a range of unintended consequences—as with any transformative new technology. Given the seemingly limitless breadth of applications for AI tools, ensuring the safe deployment of models requires, in each case, ensuring that the AI model is well suited to its operators and that the combination of the AI tool with other broader ecosystems does not generate unforeseen hazards.

AUTOMATION BIAS

The term "automation bias" refers to the tendency for individuals to excessively trust or rely on automated systems' determinations, sometimes to the detriment of performance.[93] This can occur even when the system clearly contradicts prior knowledge, intuitions, or training. In one study, for instance, participants who observed a robot perform poorly in a navigation guidance task nonetheless all chose to follow the robot minutes later in a simulated emergency evacuation, including into a dark room with no discernable exits.[94]

OPERATOR TRUST

Despite the tendency for individuals to exhibit overconfidence in automated systems' capabilities, there is also an opposite issue of ensuring that operators can maintain sufficient, appropriate trust in automated systems over time. A major emphasis of human-machine teaming research, ensuring appropriate amounts of operator trust for different types of AI-enabled systems, involves effectively communicating the systems' capabilities and limits, as well as how they perform under a range of different circumstances and with different kinds of people or teams.[95] As AI systems become more dynamic in their capabilities to execute complex tasks, the challenge of maintaining reliable, appropriate operator trust is likely to grow.[96]

Implications for Catastrophic Risk

Addressing issues associated with automation bias and operator trust is already a pressing issue in consequential systems. The U.S. Army identified automation bias as a root cause of a pair of tragic missile misfires in 2003 related to target identification system errors, resulting in the deaths of two British lieutenants and an American lieutenant in two separate incidents.[97] Though both cases are instances of friendly fire, it is not difficult to imagine a more dangerous scenario in which excess trust in a flawed automated weapons system could lead to an accidental attack on an adversary, catalyzing a cycle of rapid, violent escalation. Conversely, research by the Defense Advanced Research Projects Agency in collaboration with Marines has already highlighted the high importance of building appropriate operator trust with autonomous military systems in advance of conducting operations that would use such tools.[98]

These issues extend far beyond military systems, however. Any operator of high-risk systems could be led astray by misplaced confidence in an automated system's erroneous determination, or could fail to effectively use high-impact systems due to insufficient trust in the system, to detrimental effect. As AI systems become more capable, the temptation to be overconfident in their determinations may grow for some applications, while maintaining sufficient confidence in their capabilities may be a challenge in others.

THE LUMBERJACK EFFECT

The "lumberjack effect" suggests that the more automated a system becomes, the more difficult it is for human operators to effectively respond to system failures.[99] In other words, the higher the level of automation in a system, the harder it falls. An example of this is the 2012 Knight Capital trading accident, in which a flaw in highly automated trading software ultimately led to more than $460 million in losses to the firm.[100] Despite relatively early detection, the complexity of the system's automation meant that its technicians needed more than 20 minutes to discover how to remedy the issue, a glacial speed in the algorithmic trading world, and enough additional time for the system to make a total of four million trades at tremendous cost.[101]

ERODED SENSITIVITY TO OPERATIONS

Safety theorists have identified "sensitivity to operations" as one of five key traits that mark high reliability organizations (HROs), entities that have been remarkably effective in avoiding disasters.[102] Sensitivity to operations means that operators maintain a real-time, integrated understanding of the full breadth of complex processes they are undertaking. As a result, they are able

to quickly respond to anomalies and can more readily make sense of unexpected situations. The introduction of automation into complex processes can predictably erode this sensitivity to operations by reducing the need for operators to actively engage with and monitor the processes and environment they are overseeing.[103] This dynamic was one of the key causes of the 2009 Air France 447 tragedy, as the pilots' reliance on automated flight systems reduced their sensitivity to the flight's operations, setting the stage for the crash that killed all 228 passengers and crewmembers. In part stemming from the aircraft's automated systems, the pilots failed to fully recognize the unusual environmental conditions in which they were flying.[104]

DESKILLING AND ENFEEBLEMENT

As AI systems take over an expanding range of functions and jobs that human operators once managed, the skills needed to manage those systems can atrophy—a process known as enfeeblement or deskilling.[105] There is some precedent for this: in the aforementioned 2009 Air France 447 crash, the French Civil Aviation Safety Investigation Authority cited an erosion of flight skills related to automation as a critical factor in the crash, because the pilots had insufficient experience navigating the unusual conditions of their flight. According to analyses after the crash, the pilots would have likely gained these skills had they been trained on more flights that did not use such elaborate automation.[106]

Implications for Catastrophic Risk

The lumberjack effect, eroded sensitivity to operations, and deskilling often overlap. Each degrades operators' abilities to address problems related to AI systems as they inevitably emerge—whether because of the complexity of the automation, reduced situational awareness, or atrophied skills to accomplish the automated task manually when necessary. All three tend to take root incrementally over time, as systems become more sophisticated and further remove operators from the operational environment, and technicians' skills rust. The slow descent toward these problems makes them all the more insidious: at any one point, the extension of automation one step further in a process may make sense individually, but in aggregate can create environments of risk. Likewise, because the creep toward these issues is often slow and subtle, they are perhaps especially likely to affect high-risk systems when compared with other issues covered in this report. Whereas more obvious safety risks may receive considerable attention early on, these subtle, incremental challenges may only be noticed after it is too late. For this reason, as AI systems grow in reach and sophistication, engineers and system designers should be proactive in establishing practices

French investigators inspect debris from Air France Flight 447 for clues on the causes of the tragedy. The investigators' final report highlights both eroded sensitivity to operations and deskilling related to automation as key contributing factors. (Eric Cabanis/AFP via Getty Images)

and methods to mitigate these risks over time, including ensuring that functional analogue backup systems exist in safety-critical areas. Ensuring analogue redundancy, thereby reducing dependence on new technologies in critical processes, is a practice that has been emphasized by cybersecurity experts for several years, with clear transferability to some AI applications.[107]

TIGHT COUPLING

Tightly coupled systems are those in which constituent elements or processes within the system are directly and quickly responsive to one another, leaving little room for adjustment or flexibility. Such systems may be necessary to accomplish certain tasks requiring high efficiency, but run the risk of having cascading effects if errors cause malfunctions. Because of the close interconnectedness of tightly coupled systems, such malfunctions can be very difficult to disentangle from one another. A paradigmatic example of tight coupling is the Three Mile Island accident, in which a rapid onset of confusing, interrelated failures obscured the root causes of malfunction related to loss of coolant in a nuclear reactor, resulting in a partial meltdown. Had the failures been less closely tied to one another in a system designed to give greater room and flexibility for intervention and oversight between processes, it may have been much easier to recognize and address the core issue earlier.[108]

Implications for Catastrophic Risk

AI-powered automation can lend itself to tight coupling in systems as AI promises to speed up virtually all processes that require attention to complex details. But excess tight coupling in high-risk systems could make catastrophic events more likely across a range of domains by reducing the resiliency of these systems to errors, accelerating the impacts of errors across systems, and making the recognition of errors more difficult. Without careful attention to the dynamics of tight coupling, AI could threaten to exacerbate risks in high-impact systems across domains.

EMERGENT BEHAVIOR

Emergent behavior refers to unexpected behaviors or events that arise from the interactions between the parts of a complex system and its environment—especially if the behavior or event cannot be easily reduced to the individual effects of those parts.[109] To take a health example, if multiple medications are used to address multiple conditions in an individual, the intended effects of those medications might interact with one other in unexpected ways to produce still further effects beyond what was intended.

A classic example of AI-related emergent behavior is the case of the 2011 flash crash, in which an unknown number of lightning-speed interactions between algorithms temporarily wiped out approximately $1 trillion in stocks in a matter of minutes. Though the event is believed to have been catalyzed by misleading market behavior from one individual, the extent of the damage was caused by the complex interaction of well-functioning algorithms playing off one another in ways that were simply not anticipated.[110]

Implications for Catastrophic Risk

The integration of multiple AI tools into complex systems such as financial markets lends itself to more safety issues related to emergent behavior.[111] The current chair of the SEC has warned that the introduction of new AI tools into financial markets could lead to herding, a type of emergent behavior that can cause market instability and crashes.[112] Finance and cybersecurity are obvious candidates for emergent behavior risks, as both domains could host multiple complex systems powered by AI tools that might play off one another in unexpected ways. But other domains could also be confronted with dangerous emergent behavior if multiple AI systems interact with one another, including weapons systems.

RELEASE AND PROLIFERATION

The way that AI tools are released and proliferate can shape AI's risk profile in a range of domains. These issues have sparked considerable debate in relation to foundation models like LLMs. Advocates of open-sourcing models—making the underlying algorithms freely available—argue that open access to AI tools greatly accelerates the progress of AI research and can act as a hedge against AI companies amassing too much power as the sole proprietors of powerful tools.[113] Critics worry that such open release of AI models could pose serious risks, not least that latent, dangerous capabilities (see pages 13–15) could be exploited by bad actors in, for example, malicious hacking. In this view, once a model is open-sourced, the proliferation of that model—and its capabilities—may not be containable, and therefore providing models through "structured access" and associated safeguards is a preferable approach.[114] Proponents counter that open-source approaches to many forms of software have helped improve their security and stability and may do so in the case of AI, and that open-sourcing models may also provide incentives to develop more thorough safety mechanisms in models for open release.[115] This debate is ongoing.[116]

But the issues associated with the release and proliferation of AI tools are also broader than the debate about open-sourcing foundation models alone, with commercial and scientific incentives shaping release strategies that vary between industries and domains. For example, how specialized dual-use scientific AI tools in biology are released represents another area of concern with significant implications for risks. Additionally, the formal release strategy of an AI tool does not always determine its proliferation. Meta, for example, intended to release one of its AI models only to researchers and civil society organizations on a case-by-case basis, but the model was leaked publicly online after only a week.[117] Hackers and states may also seek to steal powerful AI tools for their own ends. These fears have been especially pronounced in relation to China, which has a track record of stealing sensitive intellectual property in an effort to catch up to and surpass the United States technologically. In March 2024, a Chinese national was charged with stealing AI research trade secrets from Google.[118]

Implications for Catastrophic Risk

AI tools with dangerous capabilities or hazardous technical deficiencies could be released in ways that drive up risks, particularly if such dangerous capabilities proliferate widely. Combined with the fact that some capabilities are latent (see pages 14–15) and that some hazards emerge only when AI tools are integrated into the broader environment, as detailed in the previous section, this heightens the relative importance of robust testing and evaluation capabilities to inform how AI tools should be appropriately released. But it also highlights the high priority of very strong security measures for developers that produce AI tools with potentially dangerous applications. Hacking groups or states such as China, Russia, Iran, or North Korea may seek to gain unauthorized access to AI tools or information with dangerous applications, routing responsible release strategies altogether—and China has already demonstrated its proclivity to do so. Theft of AI tools by such actors, with malicious intentions to use dangerous capabilities or without a thorough understanding of the dangers associated with an AI tool, could greatly exacerbate the risks of misuse and accidents for such models.

Conditions of AI Development

The conditions of AI development will inflect all the preceding dimensions of AI safety—determining the time, attention, and resources that are devoted to these issues. Though often difficult to address directly given their systemic nature, issues related to the conditions of AI development are upstream of some safety challenges, and therefore represent some of the best opportunities for early intervention as catastrophic risks continue to take shape. Though corporate and geopolitical competition is often cited as the most prominent concern for ensuring safety-friendly conditions of AI development, it is far from the only one worthy of attention. Safety cultures, investment in safety research, social resilience, and engineers' memory life cycles all play important parts in determining AI's catastrophic risk profile in the years ahead.

CORPORATE AND GEOPOLITICAL COMPETITIVE PRESSURE

As with technologies past, experts fear that competitive pressures among both AI companies and governments can quickly lead to security dilemmas and races to the bottom on safety.[119] Periods of particularly acute competition, such as heightened geopolitical tensions or aggressive commercial rivalries, can further exacerbate the issue. These pressures tend to predispose AI companies and governments alike toward pursuing speed and power over precautions and safeguards where such tradeoffs exist.[120] One example from industry is Uber's self-driving car unit, where a test vehicle struck and killed a pedestrian in 2018. Engineers disabled the vehicle's emergency braking capabilities in 2017, compelled by competitive pressures to provide a smoother rider experience.[121] In the case of states, foreign competition was a significant contributing factor to the Chernobyl meltdown: Soviet leaders selected the flawed reactor design that enabled the tragedy in part due to its distinctively Soviet development—as opposed to designs that borrowed more from American schematics—and as a quick and cost-effective option for expanding nuclear energy in the Soviet sphere of influence to keep up with the United States' ambitions to spread nuclear power in the wake of Eisenhower's "Atoms for Peace" speech.[122]

Implications for Catastrophic Risk

Escalating competitive pressures are already having effects on both corporate and state actors at the leading edge of AI development.[123] Talk of an AI race between the United States and China is now commonplace, echoing some of the dynamics that characterized technology races between great powers in the past. One danger with historical precedent

is that inaccurate perceptions of adversaries' capabilities or intentions can distort safety priorities, which is all the more relevant today given the opaque nature of assessing and verifying AI capabilities compared with conventional military hardware.[124] Appropriate caution also risks being sidelined if states fail to recognize that "superiority," conceived of only in terms of capabilities, "is not synonymous with security."[125] Narrowly focusing on who is leading in AI competition in terms of technical capacity, without accounting holistically for the risks involved in developing and deploying powerful, high-risk capabilities, can miss the forest for the trees. At the same time, the United States cannot risk falling behind its adversaries in critical areas of AI development. This is especially true in regard to China, which has a stated goal of supplanting the United States as the world leader in AI by 2030.[126] Given that companies in both countries are leading the technology's development, encouraging healthy corporate competition will likely be a strategic and economic priority for both nations, even as it can have adverse effects on safety.

Policymakers and corporate leadership must walk a fine line in ensuring that they remain competitive, but—to the extent possible—avoid the systemic safety pitfalls that often accompany competitive pressures in high-risk domains. Of course, safety and competitiveness are not always at cross-purposes, and can be mutually reinforcing.[127] But there are good reasons for which competition is often cited as a primary contributor to concerns about AI-enabled catastrophic risks. It is not difficult to imagine AI companies, with large profits on the line, cutting corners in safety to accelerate AI development for systems used in high-risk applications. Nor is it difficult to imagine countries' militaries speeding AI adoption to keep pace with one another in ways that exacerbate the risks of catastrophic accidents or miscalculations.

DEFICIENT SAFETY CULTURES

Cultures of safety vary considerably among organizations, industries, governments, and societies. Studies on HROs, for example, have demonstrated how a range of cultural traits in organizations can greatly impact the likelihood of large-scale accidents, including preoccupation with failure, reluctance to simplify, sensitivity to operations, commitment to resilience, and deference to expertise.[128] Certain industry-wide mentalities—such as those associated with Canadian-trained engineers, where initiation traditions heavily stress safety and responsibility—have also been noted to build greater safety awareness. This contrasts with other industries, for example in social media startups with their "move fast and break things" mentality that lends itself to reduced attention to safety.[129] In terms of governments, autocracies are infamous for responding poorly to budding crises, often leading to catastrophic snowball effects.[130] Finally, various societies exhibit a wide range of likelihoods for certain types of accidents and varying risk tolerances toward them, as indicated by differences in road traffic accident rates.[131] As AI is developed and deployed in diverse contexts, these differences in safety cultures will inflect the safety and stability of the resulting systems. Safety cultures that are prone to accidents are more likely to have AI-related accidents and mismanage their effects.

Implications for Catastrophic Risk

Considering the safety cultures in which AI tools are being developed and deployed is a significant, but often overlooked, priority for accurately assessing catastrophic risks associated with AI. While such an analysis is of relevance in a range of industry- and application-specific cultures, China's AI sector is particularly worthy of attention and uniquely predisposed to exacerbate catastrophic AI risks.[132] China's funding incentives around scientific and technological advancement generally lend themselves to risky approaches to new technologies, and AI leaders in China have long prided themselves on their government's large appetite for risk—even if there are more recent signs of some budding AI safety consciousness in the country.[133] China's society is the most optimistic in the world on the benefits and risks of AI technology, according to a 2022 survey by the multinational market research firm Institut Public de Sondage d'Opinion Secteur (Ipsos), despite the nation's history of grisly industrial accidents and mismanaged crises—not least its handling of COVID-19.[134] The government's sprint to lead the world in AI by 2030 has unnerving resonances with prior grand, government-led attempts to accelerate industries that have ended in tragedy, as in the Great Leap Forward, the commercial satellite launch industry, and a variety of Belt and Road infrastructure projects.[135] China's recent track record in other high-tech sectors, including space and biotech, also suggests a much greater likelihood of catastrophic outcomes.[136] Taken together, the AI-related catastrophic risks from China are particularly acute, with effects that could spread well beyond the country.

SYSTEMIC UNDERINVESTMENT IN TECHNICAL SAFETY RESEARCH AND DEVELOPMENT

Any combination of economic incentives, underestimation of risks, or misaligned interests between the

builders of AI and its users could lead to systemic underinvestment in technical AI safety capabilities relative to overall capabilities, as some argue is already the case.[137] Even though a number of leading AI labs have made safety research and development (R&D) a major priority, and the United States, United Kingdom, and Singapore have each established AI safety institutes, ensuring an appropriate balance between general AI capability research and safety research will be an ongoing challenge.[138]

Implications for Catastrophic Risk

Over time, a widening differential between general capability development in AI and technical safety development creates conditions more conducive for catastrophes. If safety capabilities fail to grow commensurately with general capabilities, the allure of integrating AI into more complex, consequential systems will climb even as the ability to ensure the trustworthiness of those systems declines. For example, if AI tools become so effective at predictive maintenance that they are increasingly relied upon for critical infrastructure management—but capabilities to ensure that their determinations are sufficiently calibrated or robust remain underdeveloped (see page 16)—critical infrastructure may become increasingly vulnerable to detrimental failures. Similarly, if vulnerabilities in AI-assisted coding tools are insufficiently recognized but are nonetheless rapidly adopted for their otherwise high-performance value, widespread coding flaws could create high-impact cyber vulnerabilities.[139]

U.S. Secretary of Commerce Gina Raimondo announced the U.S. AI Safety Institute at the AI Safety Summit at Bletchley Park on November 1, 2023. (Leon Neal via Getty Images)

SOCIAL RESILIENCE

Societies may be more or less resilient to different kinds of threats, including AI-related threats. Among the AI risks discussed in this report, social resilience has tended to be most emphasized in relation to AI-fueled disinformation. As foreign adversaries grow their information operations abroad, the degree to which disinformation campaigns are effective depends on how susceptible target populations are. Facing acute disinformation threats from China and Russia, respectively, the governments of Taiwan and Estonia have both involved their citizenries in campaigns to counter disinformation, thereby increasing the resilience of their populations.[140] Experts, including in the U.S. government, have also championed a variety of methods to improve digital literacy as a means to shore up social resilience to face increasingly sophisticated misinformation tactics, notably "prebunking" misinformation: inoculating populations to disinformation by pre-exposing them to weak forms of disinformation.[141]

But social resilience is broader than disinformation alone, and can influence the outcomes of a number of crisis scenarios. For example, the Ukrainian government's preexisting government digitization efforts inadvertently helped build its society's resistance to attacks on its critical infrastructure and government services, because citizens were already equipped with fast, adaptable systems to communicate with the government via phone apps when normal channels of communication were disrupted.[142]

Implications for Catastrophic Risk

Societies' relative resilience to degraded media ecosystems could have considerable effects on states' ability to respond to crises—influencing the degree to which smaller-scale crises can be effectively curbed from becoming full-blown catastrophes, or how great the overall impacts of a catastrophe become. As the full impacts of AI on societies' information environments take shape, promoting social resilience to misinformation or disinformation may be a critical element of mitigating catastrophic risks. Beyond AI's impacts on media ecosystems, social resilience to threats related to AI across other domains is an important feature to monitor as catastrophic AI risks continue to evolve.

ENGINEERING MEMORY LIFE CYCLES

Fading memories of the safety dynamics of certain engineering techniques over time tend to result in recurrent design faults. Sometimes these seem to occur at regular intervals: for example, in the world of large-scale bridge engineering, observers have noticed that a major collapse occurs roughly every 30 years.[143] Since this observation was made, experts have contested how regular or robust these cycles are, but have generally corroborated the underlying dynamics that drive the engineering failures in question.[144] A core feature of the issue is related to talent life cycles: generations of engineers with experience in a particular set of problems naturally phase out over time, leaving a generation that has never had to directly confront that set of issues. Additionally, slow, incremental changes in engineering techniques can cause attention to some types of failures to diminish over time, further elevating the risks of insufficient attention to seemingly dated concerns.

Implications for Catastrophic Risk

As AI development progresses at blistering speeds, ensuring appropriate attention to the full range of AI safety issues—including older, seemingly "solved" ones—will remain a challenge, especially as older generations of machine learning engineers phase out. Additionally, as AI tools are deployed into a wide range of domains, safety engineers within each must appropriately address novel safety issues associated with AI, while simultaneously not neglecting more conventional safety challenges that are often overlooked amid rapid change. Failure to do so in high-risk systems could lead to serious—if seemingly mundane—malfunctions with large-scale effects.

Further Considerations

Though necessary to explain these dimensions and issues individually, in practice it is best to consider them in combination. This is not only because the boundaries between categories often blur, but also because the historical analysis of accidents, disasters, and catastrophes suggests that crises tend to result from multiple flaws, mistakes, or errors occurring in conjunction with one another.[145] For example, simple technical failures on an AI system's embedded safety features could allow models to manifest dangerous new capabilities that its creators attempted to suppress. Some new dangerous capabilities may not emerge unless a tool is integrated into a complex ecosystem—and unknown technical faults might be compounded if a tool is integrated into a broader system.

The launches of LLMs since 2022 constitute one example of these dynamics in action on a limited scale, including OpenAI's GPT-3.5, Anthropic's Claude, Google's Bard, and Meta's Llama 2. Each company sought to embed safety features into their products to suppress the harmful content that their models could produce, but technical deficiencies in those features meant that many users were able to circumvent them to access harmful capabilities. Though companies anticipated their models' tendencies to "hallucinate" false facts, it took deployment into the broader information ecosystem to see the full effects of hallucinations. For example, after one LLM-powered chatbot hallucinated that a professor at George Washington University sexually harassed a student, another LLM-powered chatbot repeated the error, incorrectly citing as evidence an article in which the professor defended himself from the false accusation from the first chatbot.[146] Such hallucinations can be both personally damaging and, evidently, mutually reinforcing. Hypothetically, if any of these LLMs' coding abilities include unidentified technical deficiencies that produce consistent errors, their adoption among coders could introduce systemic cyber vulnerabilities, as explored in the previous section of this report, "Systemic Underinvestment in Technical Safety R&D." Already, there has been much discussion about how competition among frontier AI labs to release evermore powerful models has fueled safety compromises.[147] While illustrative, none of these recent examples comes close to catastrophic levels of impact. But other systems could—some, perhaps, in the not-too-distant future, pending when and how they are integrated into consequential systems and processes.

Perhaps the most underappreciated feature of emerging catastrophic risks related to AI from this report is the outsized likelihood of AI catastrophes originating in China. In addition to having to grapple with all the same safety challenges that other AI ecosystems must address, China's broader tech culture is prone to crisis due to its government's chronic mismanagement of disasters, censorship of information on accidents, and heavy-handed efforts to force technological breakthroughs. In AI, these dynamics are even more pronounced, buoyed by remarkably optimistic public perceptions of the technology and Beijing's gigantic strategic gamble on boosting its AI sector to international preeminence. And while both the United States and China must reckon with the safety challenges that emerge from interstate technology competitions, historically, nations that perceive themselves to be slightly behind competitors are willing to absorb the

greatest risks to catch up in tech races.[148] Thus, even while the United States' AI edge over China may be a strategic advantage, Beijing's self-perceived disadvantage could nonetheless exacerbate the overall risks of an AI catastrophe.

Failure to recognize the much more severe catastrophic risks of AI development in China could have dramatic consequences for U.S. national security. None of the most pressing catastrophic risks associated with AI discussed in this report are likely to respect national borders—whether AI-enabled bioterror pathogens spreading across the globe, AI-powered cyberattacks spinning out of control over the internet, or next-generation AI derailing intertwined financial markets. AI-enabled military decision systems going awry in China could also have very direct impacts on the United States.

Focusing risk mitigation efforts on areas where the scope for intervention is widest makes sense only to a point. Resources and thought must also be commensurate with the largest areas of risk, even if addressing those areas is far more difficult. Arguably, the discourse around managing catastrophic risks related to AI is heavily skewed toward addressing risks in open societies, despite China housing the most acute hazards. Though catastrophic risks related to AI are highly relevant across societies—and especially in the innovation-leading United States—a more holistic approach suggests that China-specific risks demand far more attention than they currently receive.

Recommendations

While a broad range of experts are working to address the issues presented in this report, there remains much more to be done, both in terms of better understanding the nature of catastrophic risks related to AI and in terms of developing viable remedies. Given the rapid development and deployment of the technology into an ever-broader range of applications, the study of catastrophic risks related to AI will also require continuous reassessment in the light of new developments. As a starting point, the following courses of action would put national security practitioners on firmer footing as they aim to address large-scale emergent risks of AI in the years ahead.

AI companies, government officials, and journalists should be more precise and deliberate in their use of terms around AI risks, particularly in reference to "catastrophic risks" and "existential risks."

Confusion between the two categories does a disservice to both and can occlude the very real catastrophic risks facing national security practitioners in the years to come. The divisive debate around existential risks need not complicate the necessary and more near-term conversation around preparing for AI-enabled catastrophic risks in national security domains. Moreover, greater attention to catastrophic risks could help better illuminate polarized debates on existential threats by showing more concretely how highly destructive events might or might not result from AI development and deployment. The UK House of Lords set an example of how to do this well in its recent report on generative AI, in which it provides clear definitions for different levels of crises, and clearly distinguishes catastrophic risks from existential risks, offering assessments for the likelihood of each separately.[149]

Building on the Biden administration's 2023 executive order on AI, relevant government agencies should more holistically explore the risks of AI integration into high-impact domains such as biosecurity, cybersecurity, finance, nuclear command and control, critical infrastructure and high-risk industries, as well as public communications.

Given the gravity of AI systems' potential impacts and current pace of progress, even seemingly small likelihoods of catastrophic events related to AI should merit sober analysis from the departments of Defense, State,

and Homeland Security. Due to the nascent stage of AI deployment in many of these areas, there continue to be "unknown unknowns" about risks, requiring capacity building and investments in resilience at multiple levels of government. As the technology continues to rapidly evolve, government assessments of catastrophic risks—beginning with those identified in the 2023 executive order—should be regularly updated, and extend beyond concerns over new capabilities and technical faults to also consider the often-overlooked dynamics of how AI's integration into complex systems and its conditions of development in different areas may contribute to safer or riskier outcomes. Given that these dimensions of catastrophic risk are often underappreciated, government agencies should ensure that AI safety research in these areas receives sufficient attention, either through funding opportunities specific to them or through explicitly requesting that projects or research into catastrophic AI risks include these elements. National security practitioners, AI labs, and industry authorities alike should pay special attention in instances where AI is deployed in new fields both inside and outside government, especially in cybersecurity, finance, nuclear command and control, and high-risk industries. Finally, the agencies and departments tasked with managing crises, such as FEMA, the Defense Threat Reduction Agency (DTRA), and the Centers for Disease Control and Prevention, as well as those tasked with examining the impacts of emerging technologies such as the Office of Science and Technology Policy, should regularly assess how evolving AI tools in the media ecosystem are changing the state's ability to respond to disasters and catastrophes (whether AI-related or not), and the public's resilience to AI-related media distortions.

Policymakers should support enhanced development of safety tests and evaluations for foundation models' capabilities.

Research into safety evaluations for foundation models is still maturing: approaches vary widely between organizations, new techniques for better eliciting capabilities are continually being discovered, and results can be sensitive to minor adjustments in methods.[150] As foundation models become more sophisticated and capable, the continuing inability to confidently ensure their performance is a long-term hazard that could exacerbate a variety of catastrophic risks associated with AI over time, especially risks related to enabling bioweapons, runaway cyberattacks, and more general loss of control issues. The ongoing debate about open-source release of models would also benefit greatly from better clarity around safety tests and evaluations for foundation models' capabilities. Policymakers should provide sustainable funding to the National Institute of Standards and Technology's U.S. AI Safety Institute, as well as to external R&D for evaluations in areas where it is apparent that the available approaches are not keeping pace with model capabilities and risks.

The departments of Defense, State, and Homeland Security should plan for AI-related catastrophes originating abroad that might impact the United States, and seek to mitigate those risks by bolstering American resilience in key domains.

Catastrophes often defy national boundaries. In the age of AI, biological, financial, military command, and cyber catastrophes could all have cascading effects well beyond an originating country's borders. The United States must be prepared for such scenarios, and help other governments understand the contributing factors to AI catastrophic risks. Teams that monitor high-risk threats abroad, such as the Department of Defense's DTRA, and the Department of State's Office of the Nonproliferation and Disarmament Fund and Office of Emerging Security Challenges, have an outsized role to play in ensuring American preparedness for catastrophic events. They must ensure that adequate information-sharing and coordination mechanisms exist among them to track catastrophic AI risks abroad. Analysis of such threats should holistically consider all the dimensions of AI catastrophic risk, including AI's integration into complex systems and the conditions of its development. Such analysis should carefully consider the cultures and incentives of technological development in other countries, most of all China, given the country's AI capabilities and unique risks.

The United States and allies must be proactive in establishing catastrophe-mitigation measures internationally where appropriate.

The United States has taken a leading role in promoting responsible AI norms in autonomous weapons, and in pushing China for clear limits on the role of AI in nuclear command.[151] However, there are more areas that would benefit from robust coordination among countries on safety measures, not least related to biosecurity. The United States should work closely with allies and partners to monitor threats, share best practices on risk reduction, and bolster cooperation in the event of AI-related catastrophes. Additionally, the United States

should recognize that building international AI safety norms requires more than safeguards and agreements, particularly as China works to build out the world's AI ecosystems concurrently with Western-led diplomatic efforts in AI safety.[152] Given the safety issues endemic to China's technological efforts, the United States will likely have to compete with China in building out AI ecosystems globally to be effective in establishing international safeguards to mitigate the risks of catastrophic outcomes.[153] American AI companies should cooperate with these efforts and explore further opportunities to collaborate internationally with peer corporations on catastrophe mitigation. To that end, industry collaboration with Chinese counterparts specifically should be a high priority. Establishing strong international industry safety norms may be one of the few viable avenues to help reduce the outsized risks of AI catastrophes originating in China, and has already shown some signs of initial progress.[154]

Conclusion

Catastrophic risks related to AI occupy a fraught position in public discourse about the technology. The historical record of international technological competition suggests that developing advanced AI capabilities is both indispensable to U.S. economic and military competitiveness and a source of potentially devastating national catastrophes. Current applications of existing advanced AI tools do not significantly alter any one catastrophic risk scenario facing the United States, but public fear about catastrophic AI risks nonetheless remains considerable—in part due to the technology's rapid advancement and the frequent conflation of "catastrophic" with "existential" risks.

Despite the confusion from these tensions in terminology, disentangling the dynamics of AI's catastrophic risks can, ultimately, be cause for optimism: the features of catastrophic risks related to AI are varied and complex, but largely manageable—as long as policymakers pay sufficient attention to all the dimensions of safety as AI systems progress. The considerable attention that has already been devoted to the issue in advance of more high-risk tools and applications suggests that policymakers are taking these risks very seriously. That being said, the most challenging risks for the foreseeable future may well be those from China's AI ecosystem, where a combination of factors makes robust AI safety far more difficult to achieve, and where the United States' influence on AI development is most limited.

For national security practitioners, catastrophic risks related to AI demand continued attention as the technology evolves, and necessitate an approach that accounts for the fact that while catastrophes are unlikely, even one would be intolerable. Keeping one step ahead of emerging risks is therefore imperative. But as they maintain a sober appreciation for the severe national security consequences of AI catastrophes, policymakers must also ensure that a fixation on worst-case scenarios does not stifle ambitions to safely realize AI's vast potential. To the contrary, ensuring resilience to AI's catastrophic risks should go hand in hand with ambitiously pursuing the immense benefits of AI for American competitiveness and security.

1. Krystal Hu, "ChatGPT Sets Record for Fastest-Growing User Base: Analyst Note," Reuters, February 2, 2023, https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01.

2. National Security Commission on Artificial Intelligence, *Final Report*, March 1, 2021, https://assets.foleon.com/eu-central-1/de-uploads-7e3kk3/48187/nscai_full_report_digital.04d6b124173c.pdf.

3. "Frontier Risk and Preparedness," OpenAI blog, October 26, 2023, https://openai.com/blog/frontier-risk-and-preparedness; "Anthropic's Responsible Scaling Policy," Anthropic, September 19, 2023, https://www.anthropic.com/news/anthropics-responsible-scaling-policy.

4. "The Bletchley Declaration by Countries Attending the AI Safety Summit, 1–2 November 2023," AI Safety Summit, November 1, 2023, https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023.

5. Anna Tong, "AI Threatens Humanity's Future, 61% of Americans Say: Reuters/Ipsos Poll," Reuters, May 17, 2023, https://www.reuters.com/technology/ai-threatens-humanitys-future-61-americans-say-reutersipsos-2023-05-17.

6. Nir Eisikovits, "AI Is an Existential Threat—Just Not the Way You Think," The Conversation, July 5, 2023, http://theconversation.com/ai-is-an-existential-threat-just-not-the-way-you-think-207680; Phil Torres, "Existential Risks: A Philosophical Analysis," *Inquiry* 66, no. 4 (April 21, 2023): 614–39, https://doi.org/10.1080/0020174X.2019.1658626; MichaelA, "Clarifying Existential Risks and Existential Catastrophes," Effective Altruism Forum, April 24, 2020, https://forum.effectivealtruism.org/posts/skPFH8LxGdKQsTkJy/clarifying-existential-risks-and-existential-catastrophes; Kamala Harris, "Remarks by Vice President Harris on the Future of Artificial Intelligence" (remarks, U.S. Embassy, London, United Kingdom, November 1, 2023), https://www.whitehouse.gov/briefing-room/speeches-remarks/2023/11/01/remarks-by-vice-president-harris-on-the-future-of-artificial-intelligence-london-united-kingdom; and *Oversight of A.I.: Rules for Artificial Intelligence: Hearing before the Senate Subcommittee on Privacy Technology and Law*, 118th Cong. (2023) (note Sen. Blumenthal's discussion of existential risk from 53:30).

7. "Pause Giant AI Experiments: An Open Letter," Future of Life Institute, March 22, 2023, https://futureoflife.org/open-letter/pause-giant-ai-experiments.

8. "Statement on AI Risk," Center for AI Safety, May 30, 2023, https://www.safe.ai/statement-on-ai-risk#signatories.

9. Eliezer Yudkowsky, "Pausing AI Developments Isn't Enough. We Need to Shut It All Down," *Time*, March 29, 2023, https://time.com/6266923/ai-eliezer-yudkowsky-open-letter-not-enough; Émile Torres, "How AI Could Accidentally Extinguish Humankind," Opinions, *The Washington Post*, August 31, 2022, https://www.washingtonpost.com/opinions/2022/08/31/artificial-intelligence-worst-case-scenario-extinction; Max Tegmark, "The 'Don't Look Up' Thinking That Could Doom Us with AI," *Time*, April 25, 2023, https://time.com/6273743/thinking-that-could-doom-us-with-ai; Ian Hogarth, "We Must Slow Down the Race to God-Like AI," *Financial Times*, April 13, 2023, https://www.ft.com/content/03895dc4-a3b7-481e-95cc-336a524f2ac2; Matthew Hutson, "Can We Stop Runaway A.I.?" *The New Yorker*, May 16, 2023, https://www.newyorker.com/science/annals-of-artificial-intelligence/can-we-stop-the-singularity.

10. "Trump Doubles Lead over DeSantis in 2024 GOP Primary Race, Quinnipiac University National Poll Finds; 65% of Voters Think Biden Is Too Old for Second Term," Quinnipiac University Poll, May 24, 2023, https://poll.qu.edu/poll-release?releaseid=3872; Harris, "Remarks by Vice President Harris on the Future of Artificial Intelligence."

11. Communications and Digital Committee, *Large Language Models and Generative AI* (London: the House of Lords, February 2, 2024), 4, https://publications.parliament.uk/pa/ld5804/ldselect/ldcomm/54/54.pdf.

12. Ann-Margaret Esnard and Alka Sapat, "Concepts and Terminology," in *Displaced by Disaster: Recovery and Resilience in a Globalizing World* (New York: Routledge, 2014).

13. Clifford E. Oliver, *Catastrophic Disaster Planning and Response*, 1st ed. (Boca Raton, FL: CRC Press, 2010), 7–8.

14. Oliver, *Catastrophic Disaster Planning and Response*, 7–8.

15. Oliver, *Catastrophic Disaster Planning and Response;* Dan Hendrycks, Mantas Mazeika, and Thomas Woodside, "An Overview of Catastrophic AI Risks" (arXiv, June 26, 2023), http://arxiv.org/abs/2306.12001; Andrew Critch and Stuart Russell, *TASRA: A Taxonomy and Analysis of Societal-Scale Risks from AI* (arXiv, June 14, 2023), https://doi.org/10.48550/arXiv.2306.06924; Nick Bostrom and Milan M. Ćirković, eds., *Global Catastrophic Risks* (Oxford: Oxford University Press, 2008).

16. Michael Coren, "Federal Agency Cites Human Error in Fatal Tesla Crash but Faults Tesla's Safeguards as 'Lacking,'" Quartz, September 12, 2017, https://qz.com/1075632/federal-agency-cites-human-error-in-fatal-tesla-tsla-crash-but-faults-teslas-safeguards-as-lacking; Jonathan M. Gitlin, "NTSB: Tesla's Autopilot UX a 'Major Role' in Fatal Model S Crash," Ars Technica, September 12, 2017, https://arstechnica.com/cars/2017/09/ntsb-teslas-autopilot-ux-a-major-role-in-fatal-model-s-crash.

17. "Deepwater Horizon: BP Gulf of Mexico Oil Spill," Environmental Protection Agency, September 12, 2013, https://www.epa.gov/enforcement/deepwater-horizon-bp-gulf-mexico-oil-spill.

18. Oliver, *Catastrophic Disaster Planning and Response*, 5.

19. "Glossary," Federal Emergency Management Agency, 2023, https://www.fema.gov/about/glossary/m.

20. Robert J. Barro, José F. Ursúa, and Joanna Weng, "The Coronavirus and the Great Influenza Pandemic: Lessons from the 'Spanish Flu' for the Coronavirus's Potential Effects on Mortality and Economic Activity," working paper, National Bureau of Economic Research, March 2020, https://doi.org/10.3386/w26866; Max Roser, "The Spanish Flu: The Global Impact of the Largest Influenza Pandemic in History," Our World in Data, July 17, 2023, https://ourworldindata.org/spanish-flu-largest-influenza-pandemic-in-history; and "The Pandemic's True Death Toll," *The Economist*, accessed April 15, 2024, https://www.economist.com/graphic-detail/coronavirus-excess-deaths-estimates.

21. Nick Bostrom, "Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards," *Journal of Evolution and Technology* 9 (2002), https://ora.ox.ac.uk/objects/uuid:827452c3-fcba-41b8-86b0-407293e6617c.

22. Bostrom, "Existential Risks."

23. Richard Ngo, Lawrence Chan, and Sören Mindermann, "The Alignment Problem from a Deep Learning Perspective" (arXiv, February 22, 2023), http://arxiv.org/abs/2209.00626; Joseph Carlsmith, "Is Power-Seeking AI an Existential Risk?" (arXiv, June 16, 2022), https://doi.org/10.48550/arXiv.2206.13353; Dan Hendrycks, "Natural Selection Favors AIs over Humans" (arXiv, May 6, 2023), https://doi.org/10.48550/arXiv.2303.16200; Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies* (Oxford: Oxford University Press, 2014); and Stuart Russell, *Human Compatible: Artificial Intelligence and the Problem of Control* (New York: Viking Press, 2019).

24. Critch and Russell, *TASRA: A Taxonomy and Analysis of Societal-Scale Risks from AI*; Andrew Critch and David Krueger, "AI Research Considerations for Human Existential Safety (ARCHES)" (arXiv, May 29, 2020), https://doi.org/10.48550/arXiv.2006.04948.

25. "Statement on AI Risk"; Yoshua Bengio et al., "Managing AI Risks in an Era of Rapid Progress" (arXiv, October 26, 2023), https://doi.org/10.48550/arXiv.2310.17688.

26. Jeremy Kahn, "Meta's A.I. Guru LeCun Wants You to Know He's No Doomer," *Fortune*, June 14, 2023, https://fortune.com/2023/06/14/metas-chief-a-i-scientist-calls-a-i-doomers-preposterous-and-predicts-llms-are-just-a-passing-fad; Brian Merchant, "Afraid of AI? The Startups Selling It Want You to Be," *Los Angeles Times*, March 31, 2023, https://www.latimes.com/business/technology/story/2023-03-31/column-afraid-of-ai-the-startups-selling-it-want-you-to-be; Timnit Gebru et al., "Statement from the Listed Authors of Stochastic Parrots on the 'AI Pause' Letter," Distributed Artificial Intelligence Research Institute, March 31, 2023, https://www.dair-institute.org/blog/letter-statement-March2023; Arvind Narayanan, Seth Lazar, and Jeremy Howard, "Is Avoiding Extinction from AI Really an Urgent Priority?" May 30, 2023, https://www.fast.ai/posts/2023-05-31-extinction.html; Jonny Thomson, "Technology Expert Tells Us Why the AI 'Doomer' Narrative Is All Wrong," Big Think, November 28, 2023, https://bigthink.com/the-future/alex-kantrowitz-why-ai-doomer-narrative-wrong; Tom Lamont, "'Humanity's Remaining Timeline? It Looks More Like Five Years than 50': Meet the Neo-Luddites Warning of an AI Apocalypse," *The Guardian*, February 17, 2024, https://www.theguardian.com/technology/2024/feb/17/humanitys-remaining-timeline-it-looks-more-like-five-years-than-50-meet-the-neo-luddites-warning-of-an-ai-apocalypse; and James Andrew Lewis, *AI and Rumors of Impending Doom* (Center for Strategic and International Studies [CSIS], July 5, 2023), https://www.csis.org/analysis/ai-and-rumors-impending-doom. See also: Kathleen Stewart and Susan Harding, "Bad Endings: American Apocalypsis," *Annual Review of Anthropology* 28, no. 1 (October 1999): 285–310, https://doi.org/10.1146/annurev.anthro.28.1.285.

27. Henry Farrell, "AI's Big Rift Is Like a Religious Schism, Says Henry Farrell," *The Economist*, December 12, 2023, https://www.economist.com/by-invitation/2023/12/12/ais-big-rift-is-like-a-religious-schism-says-henry-farrell.

28. Owen Cotton-Barratt, Max Daniel, and Anders Sandberg, "Defence in Depth against Human Extinction: Prevention, Response, Resilience, and Why They All Matter," *Global Policy* 11, no. 3 (May 2020): 271–82, https://doi.org/10.1111/1758-5899.12786.

29. See, for example: Daron Acemoglu, "Harms of AI," working paper, National Bureau of Economic Research, September 2021, https://doi.org/10.3386/w29247; Reva Schwartz et al., *Towards a Standard for Identifying and Managing Bias in Artificial Intelligence* (Gaithersburg, MD: National Institute of Standards and Technology, March 15, 2022), https://doi.org/10.6028/NIST.SP.1270; and Yuval Noah Harari, "Why Technology Favors Tyranny," *The Atlantic*, August 30, 2018, https://www.theatlantic.com/magazine/archive/2018/10/yuval-noah-harari-technology-tyranny/568330.

30. This is particularly true for the United States, where the country's geographic distance from adversaries has meant that wars spilling onto home territories are relatively rare events. But it is worth noting that new forms of interconnection could make attacks within the United States more likely. See for example, the Chinese hacking network Volt Typhoon's pre-positioned cyber assets in American critical infrastructure systems, presumably for use in the case of conflict.

31. Kenneth Payne, *I, Warbot: The Dawn of Artificially Intelligent Conflict* (New York: Oxford University Press, 2021); Paul Scharre, *Army of None: Autonomous Weap-*

ons and the Future of War (New York: W. W. Norton, 2018).

32. Colin P. Kelley et al., "Climate Change in the Fertile Crescent and Implications of the Recent Syrian Drought," *Proceedings of the National Academy of Sciences* 112, no. 11 (March 17, 2015): 3241–46, https://doi.org/10.1073/pnas.1421533112.

33. Michael Glantz and Dale Jamieson, "Societal Response to Hurricane Mitch and Intra- versus Intergenerational Equity Issues: Whose Norms Should Apply?" *Risk Analysis* 20, no. 6 (December 2000): 873–74, https://doi.org/10.1111/0272-4332.206080.

34. M. Anthony Mills, "The Masking Debate We Didn't Have," *The New Atlantis*, Summer 2023, https://www.thenewatlantis.com/publications/the-masking-debate-we-didnt-have; M. Anthony Mills, "Manufacturing Consensus," *The New Atlantis*, Fall 2021, https://www.thenewatlantis.com/publications/manufacturing-consensus; Philip Bump, "Why Do Republicans Disproportionately Believe Health Misinformation?" *The Washington Post*, August 22, 2023, https://www.washingtonpost.com/politics/2023/08/22/republicans-vaccines-polls; and Charlotte Klein, "Right-Wing Vaccine Lies Are Tearing the Country Apart," *Vanity Fair*, July 15, 2021, https://www.vanityfair.com/news/2021/07/right-wing-vaccine-lies-are-tearing-the-country-apart.

35. Elizabeth Seger, "Should Epistemic Security Be a Priority GCR Cause Area?" in *Intersections, Reinforcements, Cascades: Proceedings of the 2023 Stanford Existential Risks Conference* (Stanford Digital Repository, 2023), https://doi.org/10.25740/bc884qy3778.

36. Frank Dikötter, *Mao's Great Famine: The History of China's Most Devastating Catastrophe, 1958–1962* (New York: Walker Books, 2010).

37. Miguel Centeno et al., eds., *How Worlds Collapse* (New York: Routledge, 2023), 30–31.

38. Mark Pelling and Kathleen Dill, "Disaster Politics: Tipping Points for Change in the Adaptation of Sociopolitical Regimes," *Progress in Human Geography* 34 (February 1, 2010): 24, https://doi.org/10.1177/0309132509105004.

39. "Normalcy Bias," The Decision Lab, accessed January 22, 2024, https://thedecisionlab.com/biases/normalcy-bias; Nassim Nicholas Taleb, *The Black Swan: The Impact of the Highly Improbable* (New York: Penguin, 2008).

40. International Atomic Energy Agency, *Ten Years after Chernobyl: What Do We Really Know?* (Vienna, Austria: Division of Public Information, 1996), 8, https://inis.iaea.org/collection/NCLCollectionStore/_Public/28/058/28058918.pdf; Mikhail Gorbachev, "Turning Point at Chernobyl," *Project Syndicate*, April 14, 2006, https://www.project-syndicate.org/commentary/turning-point-at-chernobyl; Mark Joseph Stern, "Did Chernobyl Cause the Soviet Union To Explode?" *Slate*, January 25, 2013, https://slate.com/technology/2013/01/chernobyl-and-the-fall-of-the-soviet-union-gorbachevs-glasnost-allowed-the-nuclear-catastrophe-to-undermine-the-ussr.html.

41. Toby Ord, "Anthropogenic Risks" and "Future Risks," in *The Precipice: Existential Risk and the Future of Humanity* (New York: Hachette Books, 2020), 89–158.

42. This figure is adapted from a graph found in Dupuy's *The Evolution of Weapons and Warfare*, building on earlier work in *Numbers, Predictions, and War,* in which he charts values that "represent [weapons'] destructive force in terms of the number of men a weapon can theoretically kill in one hour under certain artificial, laboratory-like circumstances." Trevor N. Dupuy, *The Evolution of Weapons And Warfare* (New York: Da Capo Press, 1990), 92, 286–94; *Numbers, Predictions, and War: Using History to Evaluate Combat Factors and Predict the Outcome of Battles* (Indianapolis: Bobbs-Merill, 1979), 6.

43. Alexander Kott, "Toward Universal Laws of Technology Evolution: Modeling Multi-Century Advances in Mobile Direct-Fire Systems," *The Journal of Defense Modeling and Simulation* 17, no. 4 (October 1, 2020): 373–88, https://doi.org/10.1177/1548512919875523.

44. Tanisha M. Fazal and Paul Poast, "War Is Not Over," *Foreign Affairs*, October 15, 2019, https://www.foreignaffairs.com/world/war-not-over; Dupuy, *The Evolution of Weapons and Warfare*, 92, 286–94.

45. See, for example, Jonathan Glancey, "Crashes That Changed Plane Design," BBC, February 24, 2022, https://www.bbc.com/future/article/20140414-crashes-that-changed-plane-design; Tim Fernholz, "U.S. Nuclear Tests Killed Far More Civilians Than We Knew," Quartz, December 21, 2017, https://qz.com/1163140/us-nuclear-tests-killed-american-civilians-on-a-scale-comparable-to-hiroshima-and-nagasaki.

46. "Annual Global Corporate Investment in Artificial Intelligence, by Type," Our World in Data, June 14, 2023, https://ourworldindata.org/grapher/corporate-investment-in-artificial-intelligence-by-type.

47. Fabio Urbina et al., "Dual Use of Artificial Intelligence-Powered Drug Discovery," *Nature Machine Intelligence* 4, no. 3 (March 2022): 189–91, https://doi.org/10.1038/s42256-022-00465-9; Emily H. Soice et al., "Can Large Language Models Democratize Access to Dual-Use Biotechnology?" (arXiv, June 6, 2023), https://doi.org/10.48550/arXiv.2306.03809; Robert Service, "Could Chatbots Help Devise the next Pandemic Virus?" *Science Insider*, June 14, 2023, https://www.science.org/content/article/could-chatbots-help-devise-next-pandemic-virus; Daniil A. Boiko, Robert MacKnight, and Gabe Gomes, "Emergent Autonomous Scientific Research Capabilities of Large Language Models" (arXiv, April 11, 2023), http://arxiv.org/abs/2304.05332; "Frontier Threats Red Teaming for AI Safety," Anthropic blog, July 26, 2023, https://www.anthropic.com/index/frontier-threats-red-teaming-for-ai-safety; and Christopher A. Mouton, Caleb Lucas, and Ella Guest, *The Operational Risks of AI in Large-Scale Biological Attacks: A Red-Team Approach* (Santa

Monica, CA: RAND Corporation, October 16, 2023), https://www.rand.org/pubs/research_reports/RRA2977-1.html.

48. Dionysios Demetis and Allen Lee, "When Humans Using the IT Artifact Becomes IT Using the Human Artifact," *Journal of the Association for Information Systems* 19, no. 10 (October 31, 2018), https://aisel.aisnet.org/jais/vol19/iss10/5.

49. Andrew Ross Sorkin, "The S.E.C.'s Gensler Is Worried about A.I.," DealBook, *The New York Times*, August 7, 2023, https://messaging-custom-newsletters.nytimes.com/template/oakv2?campaign_id=4&emc=edit_dk_20230807&instance_id=99445&nl=dealbook&productCode=DK&regi_id=56869983&segment_id=141330&te=1&uri=nyt%3A%2F%2Fnewsletter%2F2fbfe325-70de-5c7a-9a74-62753eda29d5.

50. Simon Shuster, "Stanislav Petrov, the Russian Officer Who Averted a Nuclear War, Feared History Repeating Itself," *Time*, September 19, 2017, https://time.com/4947879/stanislav-petrov-russia-nuclear-war-obituary; Patricia Lewis et al., *Too Close for Comfort: Cases of Near Nuclear Use and Options for Policy* (London: Chatham House, the Royal Institute of International Affairs, April 2014), https://www.chathamhouse.org/sites/default/files/field/field_document/20140428TooCloseforComfortNuclearUseLewisWilliamsPelopidasAghlani.pdf; Anthony Aguirre, Emilia Javorsky, and Max Tegmark, "'Artificial Escalation': Imagining the Future of Nuclear Risk," Bulletin of the Atomic Scientists, July 17, 2023, https://thebulletin.org/2023/07/artificial-escalation-imagining-the-future-of-nuclear-risk.

51. Ted Lieu, "Reps. Lieu, Buck, Eshoo and Sen. Schatz Introduce Bipartisan, Bicameral Bill to Create a National Commission on Artificial Intelligence," press release, June 20, 2023, http://lieu.house.gov/media-center/press-releases/reps-lieu-buck-eshoo-and-sen-schatz-introduce-bipartisan-bicameral-bill; "Majority Leader Schumer Delivers Remarks to Launch SAFE Innovation Framework for Artificial Intelligence at CSIS" (speech, CSIS and Senate Democrats, Washington, DC, June 21, 2023), https://www.democrats.senate.gov/news/press-releases/majority-leader-schumer-delivers-remarks-to-launch-safe-innovation-framework-for-artificial-intelligence-at-csis; Cecilia Kang, "OpenAI's Sam Altman Urges A.I. Regulation in Senate Hearing," *The New York Times*, May 16, 2023, https://www.nytimes.com/2023/05/16/technology/openai-altman-artificial-intelligence-regulation.html; Ashley Gold and Maria Curi, "Scoop: Thune Readies AI Certification Bill," Axios, July 18, 2023, https://www.axios.com/pro/tech-policy/2023/07/18/thune-readies-ai-certification-bill; and The White House, "Fact Sheet: Biden-Harris Administration Secures Voluntary Commitments from Leading Artificial Intelligence Companies to Manage the Risks Posed by AI," press release, July 21, 2023, https://www.whitehouse.gov/briefing-room/statements-releases/2023/07/21/fact-sheet-biden-harris-administration-secures-voluntary-commitments-from-leading-artificial-intelligence-companies-to-manage-the-risks-posed-by-ai.

52. John T. O'Brien and Cassidy Nelson, "Assessing the Risks Posed by the Convergence of Artificial Intelligence and Biotechnology," *Health Security* 18, no. 3 (June 1, 2020): 219–27, https://doi.org/10.1089/hs.2019.0122.

53. Seger, "Should Epistemic Security Be a Priority GCR Cause Area?"; Elizabeth Seger et al., *Tackling Threats to Informed Decisionmaking in Democratic Societies: Promoting Epistemic Security in a Technologically-Advanced World* (London: The Alan Turing Institute, October 2020), https://www.turing.ac.uk/news/publications/tackling-threats-informed-decision-making-democratic-societies.

54. Josh A. Goldstein et al., "Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations" (arXiv, January 10, 2023), http://arxiv.org/abs/2301.04246; Canyu Chen and Kai Shu, "Combating Misinformation in the Age of LLMs: Opportunities and Challenges" (arXiv, November 8, 2023), https://doi.org/10.48550/arXiv.2311.05656; Ali Swenson and Will Weissert, "New Hampshire Investigating Fake Biden Robocall Meant to Discourage Voters Ahead of Primary," AP News, January 22, 2024, https://apnews.com/article/new-hampshire-primary-biden-ai-deepfake-robocall-f3469ceb6dd613079092287994663db5; and Jane Wakefield, "Deepfake Presidents Used in Russia-Ukraine War," BBC, March 18, 2022, https://www.bbc.com/news/technology-60780142.

55. Brian Klaas, "Vladimir Putin Has Fallen into the Dictator Trap," *The Atlantic*, March 16, 2022, https://www.theatlantic.com/ideas/archive/2022/03/putin-dictator-trap-russia-ukraine/627064; Susan Shirk, "China in Xi's 'New Era': The Return to Personalistic Rule," *Journal of Democracy* 29, no. 2 (April 2018): 22–36.

56. See: Dikötter, *Mao's Great Famine*.

57. Dario Amodei et al., "Concrete Problems in AI Safety" (arXiv, July 25, 2016), https://doi.org/10.48550/arXiv.1606.06565; Dan Hendrycks et al., "Unsolved Problems in ML Safety" (arXiv, June 16, 2022), http://arxiv.org/abs/2109.13916; Hendrycks, Mazeika, and Woodside, "An Overview of Catastrophic AI Risks"; Critch and Russell, *TASRA: A Taxonomy and Analysis of Societal-Scale Risks from AI*.

58. Will Hunt, "*The Flight to Safety-Critical AI: Lessons in AI Safety from the Aviation Industry*" (Berkeley, CA: UC Berkeley Center for Long-Term Cybersecurity, August 2020), https://cltc.berkeley.edu/wp-content/uploads/2020/08/Flight-to-Safety-Critical-AI.pdf.

59. Taleb, *The Black Swan*, chap. 1.

60. Remco Zwetsloot and Allan Dafoe, "Thinking about Risks from AI: Accidents, Misuse and Structure," Lawfare, February 11, 2019, https://www.lawfaremedia.org/article/thinking-about-risks-ai-accidents-misuse-and-structure.

61. Graham Webster et al., "Full Translation: China's 'New

Generation Artificial Intelligence Development Plan,'" New America, August 1, 2017, http://newamerica.org/cybersecurity-initiative/digichina/blog/full-translation-chinas-new-generation-artificial-intelligence-development-plan-2017.

62. Bill Drexel and Hannah Kelley, "China Is Flirting with AI Catastrophe," *Foreign Affairs*, May 30, 2023, https://www.foreignaffairs.com/china/china-flirting-ai-catastrophe. For a more comprehensive exploration of the military and economic stakes of Sino-American AI competition, see Bill Drexel and Hannah Kelley, "Behind China's Plans to Build AI for the World," *POLITICO*, November 30, 2023, https://www.politico.com/news/magazine/2023/11/30/china-global-ai-plans-00129160.

63. Goldstein et al., "Generative Language Models and Automated Influence Operations."

64. Stephanie Batalis, "AI and Biorisk: An Explainer," (Washington, DC: Center for Security and Emerging Technology [CSET], December 2023), https://cset.georgetown.edu/publication/ai-and-biorisk-an-explainer; Urbina et al., "Dual Use of Artificial Intelligence-Powered Drug Discovery"; Jonas B. Sandbrink, "Artificial Intelligence and Biological Misuse: Differentiating Risks of Language Models and Biological Design Tools" (arXiv, August 12, 2023), http://arxiv.org/abs/2306.13952; and Tejal Patwardhan et al., *Building an Early Warning System for LLM-Aided Biological Threat Creation* (OpenAI, January 31, 2024), https://openai.com/research/building-an-early-warning-system-for-llm-aided-biological-threat-creation.

65. Bob Violino, "AI Tools Such as ChatGPT Are Generating a Mammoth Increase in Malicious Phishing Emails," CNBC, November 28, 2023, https://www.cnbc.com/2023/11/28/ai-like-chatgpt-is-creating-huge-increase-in-malicious-phishing-email.html; "The Near-Term Impact of AI on the Cyber Threat," National Cyber Security Centre (United Kingdom), January 24, 2024, https://www.ncsc.gov.uk/report/impact-of-ai-on-cyber-threat; Andrew Lohn and Krystal Jackson, *Will AI Make Cyber Swords or Shields?* (Washington, DC: CSET, August 2022), https://cset.georgetown.edu/publication/will-ai-make-cyber-swords-or-shields; and Ben Buchanan et al., *Automating Cyber Attacks: Hype and Reality* (Washington, DC: CSET, November 2020), https://cset.georgetown.edu/publication/automating-cyber-attacks.

66. Batalis, *AI and Biorisk*; Urbina et al., "Dual Use of Artificial Intelligence-Powered Drug Discovery"; Jonas B. Sandbrink, "Artificial Intelligence and Biological Misuse: Differentiating Risks of Language Models and Biological Design Tools" (arXiv, August 12, 2023), http://arxiv.org/abs/2306.13952; and Patwardhan et al., *Building an Early Warning System for LLM-Aided Biological Threat Creation*.

67. "The Near-Term Impact of AI on the Cyber Threat"; Lohn and Jackson, *Will AI Make Cyber Swords or Shields?*; and Buchanan et al., *Automating Cyber Attacks*.

68. Seger, "Should Epistemic Security Be a Priority GCR Cause Area?"; Seger et al., *Tackling Threats to Informed Decisionmaking in Democratic Societies*.

69. James Johnson, *AI and the Bomb: Nuclear Strategy and Risk in the Digital Age* (Oxford: Oxford University Press, 2023).

70. The empirical observation that models leveraging the most computation are the most capable is sometimes known as "The Bitter Lesson," a term popularized by AI research scientist Rich Sutton, in: Rich Sutton, "The Bitter Lesson," Incomplete Ideas, March 13, 2019, http://www.incompleteideas.net/IncIdeas/BitterLesson.html. This observation has been characterized in "scaling laws," which describe how model capabilities scale with training inputs: Pablo Villalobos, *Scaling Laws Literature Review*, Epoch AI, January 26, 2023, https://epochai.org/blog/scaling-laws-literature-review.

71. Jason Wei et al., "Emergent Abilities of Large Language Models" (arXiv, October 26, 2022), https://doi.org/10.48550/arXiv.2206.07682. Claims of emergent capabilities have been questioned, with arguments that they are simply artifacts of how performance was measured: Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo, "Are Emergent Abilities of Large Language Models a Mirage?" (arXiv, May 22, 2023), https://doi.org/10.48550/arXiv.2304.15004. Wei has responded (including examples of emergent abilities that are robust to this critique): Wei, "Common Arguments Regarding Emergent Abilities," Jason Wei blog, May 3, 2023, https://www.jasonwei.net/blog/common-arguments-regarding-emergent-abilities. For further discussion, see Thomas Woodside, "Emergent Abilities in Large Language Models: An Explainer," CSET, April 16, 2024, https://cset.georgetown.edu/article/emergent-abilities-in-large-language-models-an-explainer; Markus Anderljung et al., "Appendix B: Scaling laws in Deep Learning," in *Frontier AI Regulation: Managing Emerging Risks to Public Safety* (arXiv, September 4, 2023), https://doi.org/10.48550/arXiv.2307.03718.

72. Hendrycks et al., "Unsolved Problems in ML Safety," 7.

73. Anderljung et al., "Frontier AI Regulation," sec. 2.2.1.

74. Tom Davidson et al., "AI Capabilities Can Be Significantly Improved without Expensive Retraining" (arXiv, December 12, 2023), https://doi.org/10.48550/arXiv.2312.07413.

75. Kevin Lu et al., "Pretrained Transformers as Universal Computation Engines" (arXiv, June 30, 2021), https://doi.org/10.48550/arXiv.2103.05247.

76. Kuang-Huei Lee et al., "Multi-Game Decision Transformers," in *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*, 36th Conference on Neural Information Processing Systems, New Orleans, Louisiana, 2022, https://proceedings.neurips.cc/paper_files/pa-

per/2022/hash/b2cac94f82928a85055987d9fd44753f-Abstract-Conference.html; Jindong Wang et al., "Generalizing to Unseen Domains: A Survey on Domain Generalization," *IEEE Transactions on Knowledge and Data Engineering* 35, no. 8 (August 2023): 8052–72, https://doi.org/10.1109/TKDE.2022.3178128.

77. Urbina et al., "Dual Use of Artificial Intelligence-Powered Drug Discovery."

78. Victoria Krakovna et al., "Specification Gaming: The Flip Side of AI Ingenuity," Google DeepMind blog, April 21, 2020, https://deepmind.google/discover/blog/specification-gaming-the-flip-side-of-ai-ingenuity; Tim G. J. Rudner and Helen Toner, *Key Concepts in AI Safety: Specification in Machine Learning* (Washington, DC: CSET, December 2021), https://cset.georgetown.edu/publication/key-concepts-in-ai-safety-specification-in-machine-learning; Hendrycks et al., "Unsolved Problems in ML Safety," sec. 4; Hendrycks, Mazeika, and Woodside, "An Overview of Catastrophic AI Risks," sec. 5.1; and Alexander Pan, Kush Bhatia, and Jacob Steinhardt, "The Effects of Reward Misspecification: Mapping and Mitigating Misaligned Models" (arXiv, February 14, 2022), https://doi.org/10.48550/arXiv.2201.03544.

79. Jack Clark and Dario Amodei, "Faulty Reward Functions in the Wild," OpenAI blog, December 21, 2016, https://openai.com/research/faulty-reward-functions.

80. Hendrycks, Mazeika, and Woodside, "An Overview of Catastrophic AI Risks," 34–35.

81. Andy Greenberg, "The Untold Story of NotPetya, the Most Devastating Cyberattack in History," *Wired*, August 22, 2018, https://www.wired.com/story/notpetya-cyberattack-ukraine-russia-code-crashed-the-world.

82. Rudner and Toner, *Key Concepts in AI Safety: Specification in Machine Learning;* Jacob Steinhardt and Helen Toner, "Why Robustness Is Key to Deploying AI," The Brookings Institution, June 8, 2020, https://www.brookings.edu/articles/why-robustness-is-key-to-deploying-ai; Hendrycks et al., "Unsolved Problems in ML Safety," secs. 2.1, 3.1.

83. Rohin Shah et al., "Goal Misgeneralization: Why Correct Specifications Aren't Enough for Correct Goals" (arXiv, November 2, 2022), https://doi.org/10.48550/arXiv.2210.01790.

84. Hendrycks et al., "Unsolved Problems in ML Safety," sec. 3.2.1; Matthias Minderer et al., "Revisiting the Calibration of Modern Neural Networks" (arXiv, October 26, 2021), https://doi.org/10.48550/arXiv.2106.07998; Chuan Guo et al., "On Calibration of Modern Neural Networks," in *Proceedings of the 34th International Conference on Machine Learning, Volume 70*, ICML'17 (Sydney, Australia: JMLR.org, 2017), 1321–30, https://proceedings.mlr.press/v70/guo17a/guo17a.pdf.

85. Anthony Corso et al., "A Holistic Assessment of the Reliability of Machine Learning Systems" (arXiv, July 29, 2023), 16–18, https://doi.org/10.48550/arXiv.2307.10586.

86. Apostol Vassilev et al., *Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations* (Gaithersburg, MD: National Institute of Standards and Technology, January 2024), https://doi.org/10.6028/NIST.AI.100-2e2023; Tim G. J. Rudner and Helen Toner, *Key Concepts in AI Safety: Robustness and Adversarial Examples* (Washington, DC: CSET, March 2021), https://cset.georgetown.edu/publication/key-concepts-in-ai-safety-robustness-and-adversarial-examples; and Hendrycks et al., "Unsolved Problems in ML Safety," secs. 2.2, 3.3.1.

87. Paul Scharre, "Poison," in *Four Battlegrounds: Power in the Age of Artificial Intelligence* (W. W. Norton, 2024).

88. Pengfei Jing et al., "Too Good to Be Safe: Tricking Lane Detection in Autonomous Driving with Crafted Perturbations" (30th USENIX Security Symposium, virtual event, 2021), 3237–54, https://www.usenix.org/conference/usenixsecurity21/presentation/jing; Mahmood Sharif et al., "Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, CCS '16 (New York: Association for Computing Machinery, 2016), 1528–40, https://doi.org/10.1145/2976749.2978392.

89. The neurons in neural networks are inspired by biological neurons, but with important differences.

90. Wojciech Samek, Thomas Wiegand, and Klaus-Robert Müller, "Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models," *ITU Journal: ICT Discoveries* 1, no. 1 (October 13, 2017): 39–48, https://itu-ilibrary.org/science-and-technology/explainable-artificial-intelligence_pub/8129fdff-en; Tim G. J. Rudner and Helen Toner, *Key Concepts in AI Safety: Interpretability in Machine Learning* (Washington, DC: CSET, March 2021), https://cset.georgetown.edu/publication/key-concepts-in-ai-safety-interpretability-in-machine-learning; Rishi Bommasani et al., "On the Opportunities and Risks of Foundation Models" (arXiv, July 12, 2022), sec. 4.11, https://doi.org/10.48550/arXiv.2108.07258; Elham Tabassi, *Artificial Intelligence Risk Management Framework (AI RMF 1.0)* (Gaithersburg, MD: National Institute of Standards and Technology, January 26, 2023), sec. 3.5, https://doi.org/10.6028/NIST.AI.100-1; and Tilman Räuker et al., "Toward Transparent AI: A Survey on Interpreting the Inner Structures of Deep Neural Networks" (arXiv.org, July 27, 2022), https://arxiv.org/abs/2207.13243v6.

91. Jonathan Zittrain, "The Hidden Costs of Automated Thinking," *The New Yorker*, July 23, 2019, https://www.newyorker.com/tech/annals-of-technology/the-hidden-costs-of-automated-thinking.

92. D. M. Murphy and M. E. Paté-Cornell, "The SAM Framework: Modeling the Effects of Management

Factors on Human Behavior in Risk Analysis," *Risk Analysis* 16, no. 4 (August 1996): 501–15, https://doi.org/10.1111/j.1539-6924.1996.tb01096.x.

93. Kate Goddard, Abdul Roudsari, and Jeremy C Wyatt, "Automation Bias: A Systematic Review of Frequency, Effect Mediators, and Mitigators," *Journal of the American Medical Informatics Association* 19, no. 1 (2012): 121–27, https://doi.org/10.1136/amiajnl-2011-000089; Linda J. Skitka, Kathleen L. Mosier, and Mark Burdick, "Does Automation Bias Decision-Making?" *International Journal of Human-Computer Studies* 51, no. 5 (November 1999): 991–1006, https://doi.org/10.1006/ijhc.1999.0252.

94. Paul Robinette et al., "Overtrust of Robots in Emergency Evacuation Scenarios," in *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2016, 101–8, https://doi.org/10.1109/HRI.2016.7451740.

95. Dewey Murdick, *Building Trust in AI: A New Era of Human-Machine Teaming* (Washington, DC: CSET, July 19, 2023), https://cset.georgetown.edu/article/building-trust-in-ai-a-new-era-of-human-machine-teaming.

96. Heather M. Roff and David Danks, "'Trust but Verify': The Difficulty of Trusting Autonomous Weapons Systems," *Journal of Military Ethics* 17, no. 1 (January 2, 2018): 2–20, https://doi.org/10.1080/15027570.2018.1481907.

97. Scharre, *Army of None*, chap. 9.

98. Adam Stone, "Knock, Knock. Who's There? This AI Combat System Might Already Know," C4ISRNet, January 31, 2019, https://www.c4isrnet.com/it-networks/2019/01/31/knock-knock-whos-there-this-ai-combat-system-might-already-know.

99. Allyson I. Hauptman and Nathan J. McNeese, "Overcoming the Lumberjack Effect through Adaptive Autonomy," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 66, no. 1 (September 1, 2022): 1075–79, https://doi.org/10.1177/1071181322661372.

100. Securities and Exchange Commission, "In the Matter of Knight Capital Americas LLC," Release No. 70694, File No. 3-15570 (October 16, 2013), https://www.sec.gov/files/litigation/admin/2013/34-70694.pdf.

101. "Case Study 4: The $440 Million Software Error at Knight Capital," Henrico Dolfing, June 5, 2019, https://www.henricodolfing.com/2019/06/project-failure-case-study-knight-capital.html; Jessica Silver-Greenberg, Nathaniel Popper, and Michael De La Merced, "Trading Program Ran Amok, with No 'Off' Switch," Dealbook, *The New York Times*, August 3, 2012, https://dealbook.nytimes.com/2012/08/03/trading-program-ran-amok-with-no-off-switch.

102. See Karl E. Weick and Kathleen M. Sutcliffe, *Managing the Unexpected: Sustained Performance in a Complex World*, 3rd ed. (Hoboken, New Jersey: Jossey-Bass, 2015).

103. Nick Oliver, Thomas Calvard, and Kristina Potočnik, "Cognition, Technology, and Organizational Limits: Lessons from the Air France 447 Disaster," *Organization Science* 28, no. 4 (August 2017): 729–43, https://doi.org/10.1287/orsc.2017.1138.

104. Eric Marsden, "Air France Flight 447: Confusion on the Flight Deck," Risk Engineering, April 5, 2017, https://risk-engineering.org/concept/AF447-Rio-Paris.

105. Critch and Krueger, "AI Research Considerations for Human Existential Safety (ARCHES)," 31; Hendrycks, Mazeika, and Woodside, "An Overview of Catastrophic AI Risks," 19.

106. "Final Report on the Accident on 1st June 2009 to the Airbus A330-203 Registered F-GZCP Operated by Air France Flight AF 447 Rio de Janeiro – Paris" (Paris: Bureau d'Enquêtes et d'Analyses pour la sécurité de l'aviation civile, July 2012), 185–86, https://bea.aero/fileadmin/documents/docspa/2009/f-cp090601.en/pdf/f-cp090601.en.pdf; William Langewiesche, "Should Airplanes Be Flying Themselves?" *Vanity Fair*, September 17, 2014, https://www.vanityfair.com/news/business/2014/10/air-france-flight-447-crash.

107. David Sax, "In the Age of Cybercrime, the Best Insurance May Be Analog," *Bloomberg*, March 10, 2016, https://www.bloomberg.com/news/articles/2016-03-10/cybersecurity-the-best-insurance-may-be-analog.

108. Charles Perrow, *Normal Accidents: Living with High-Risk Technologies* (Princeton, NJ: Princeton University Press, 1999), chap. 3.

109. Chris Johnson, "What Are Emergent Properties and How Do They Affect the Engineering of Complex Systems?" *Reliability Engineering & System Safety* 91 (December 1, 2006), https://doi.org/10.1016/j.ress.2006.01.008.

110. Demetis and Lee, "When Humans Using the IT Artifact Becomes IT Using the Human Artifact."

111. Lorenzo Barberis Canonico and Nathan McNeese, "Flash Crashes in Multi-Agent Systems Using Minority Games and Reinforcement Learning to Test AI Safety," in *2019 Winter Simulation Conference (WSC)*, 2019, 193–204, https://doi.org/10.1109/WSC40007.2019.9004675.

112. Gary Gensler and Lily Bailey, "Deep Learning and Financial Stability," working paper, Artificial Intelligence eJournal, Social Science Research Network (November 1, 2020), https://doi.org/10.2139/ssrn.3723132.

113. Jessica Billingsley, "Beyond Corporate AI: Why We Need an Open-Source Revolution," *Rolling Stone*, November 1, 2023, https://www.rollingstone.com/culture-council/articles/beyond-corporate-ai-why-we-need-open-source-revolution-1234867419.

114. Elizabeth Seger et al., "Open-Sourcing Highly Capable Foundation Models: An Evaluation of Risks, Benefits, and Alternative Methods for Pursuing Open-Source Objectives" (arXiv, September 29, 2023), https://doi.org/10.48550/arXiv.2311.09227.

115. Rahul Roy-Chowdhury, "Why Open-Source Is Crucial for Responsible AI Development," World Economic Forum, December 22, 2023, https://www.weforum.org/agenda/2023/12/ai-regulation-open-source.

116. Kyle Miller, "Open Foundation Models: Implications of Contemporary Artificial Intelligence," CSET, March 12, 2024, https://cset.georgetown.edu/article/open-foundation-models-implications-of-contemporary-artificial-intelligence; Caleb Withers, *Response to NTIA Request for Comment: "Dual Use Foundation Artificial Intelligence Models with Widely Available Model Weights"* (Washington, DC: Center for a New American Security, March 27, 2024), secs. 3., 6.a., https://www.cnas.org/publications/commentary/response-to-ntia-request-for-comment-dual-use-foundation-artificial-intelligence-models-with-widely-available-model-weights.

117. James Vincent, "Meta's Powerful AI Language Model Has Leaked Online—What Happens Now?" The Verge, March 8, 2023, https://www.theverge.com/2023/3/8/23629362/meta-ai-language-model-llama-leak-online-misuse.

118. Dustin Volz and Robert McMillan, "U.S. Charges Chinese National with Stealing AI Secrets from Google," *The Wall Street Journal*, March 6, 2024, https://www.wsj.com/politics/national-security/u-s-charges-chinese-national-with-stealing-ai-secrets-from-google-5c66524a.

119. Paul Scharre, "Debunking the AI Arms Race Theory," *Texas National Security Review* 4, no. 3 (June 28, 2021): 121–132, https://tnsr.org/2021/06/debunking-the-ai-arms-race-theory.

120. Stuart Armstrong, Nick Bostrom, and Carl Shulman, "Racing to the Precipice: A Model of Artificial Intelligence Development," *AI & SOCIETY* 31, no. 2 (May 1, 2016): 201–6, https://doi.org/10.1007/s00146-015-0590-y.

121. Julie Bort, "Uber Insiders Describe Infighting and Questionable Decisions before Its Self-Driving Car Killed a Pedestrian," Business Insider, November 19, 2018, https://www.businessinsider.com/sources-describe-questionable-decisions-and-dysfunction-inside-ubers-self-driving-unit-before-one-of-its-cars-killed-a-pedestrian-2018-10.

122. Sonja D. Schmid, *Producing Power: The Pre-Chernobyl History of the Soviet Nuclear Industry* (Cambridge, MA: MIT Press, 2015), 101, 112, 121–23.

123. Nico Grant, "Google Calls In Help from Larry Page and Sergey Brin for A.I. Fight," *The New York Times*, January 20, 2023, https://www.nytimes.com/2023/01/20/technology/google-chatgpt-artificial-intelligence.html; Dylan Matthews, "The $1 Billion Gamble to Ensure AI Doesn't Destroy Humanity," Vox, July 17, 2023, https://www.vox.com/future-perfect/23794855/anthropic-ai-openai-claude-2.

124. Mariano-Florentino Cuéllar and Matt Sheehan, "AI Is Winning the AI Race," *Foreign Policy*, June 19, 2023, https://foreignpolicy.com/2023/06/19/us-china-ai-race-regulation-artificial-intelligence; Helen Toner, Jenny Xiao, and Jeffrey Ding, "The Illusion of China's AI Prowess," *Foreign Affairs*, June 2, 2023, https://www.foreignaffairs.com/china/illusion-chinas-ai-prowess-regulation; Jane Vaynman, "Better Monitoring and Better Spying: The Implications of Emerging Technology for Arms Control," *Texas National Security Review* 4, no. 4 (Fall 2021): 33–56, https://tnsr.org/2021/09/better-monitoring-and-better-spying-the-implications-of-emerging-technology-for-arms-control.

125. Richard Danzig, *Technology Roulette: Managing Loss of Control as Many Militaries Pursue Technological Superiority* (Washington, DC: Center for a New American Security, May 30, 2018), https://www.cnas.org/publications/reports/technology-roulette.

126. Webster et al., "Full Translation: China's 'New Generation Artificial Intelligence Development Plan.'"

127. Steven Schavell and A. Mitchell Polinsky, "The Uneasy Case for Product Liability," *Harvard Law Review* 123, no. 6 (April 20, 2010), https://harvardlawreview.org/print/vol-123/the-uneasy-case-for-product-liability; Andrew F. Daughety and Jennifer F. Reinganum, "Product Safety: Liability, R&D, and Signaling," *The American Economic Review* 85, no. 5 (1995): 1187–1206, https://www.jstor.org/stable/2950983.

128. Weick and Sutcliffe, *Managing the Unexpected*.

129. Henry Petroski, *To Forgive Design: Understanding Failure* (Cambridge, MA: Belknap Press of Harvard University Press, 2014), chap. 8.

130. Frances D'Souza, "Democracy as a Cure for Famine," *Journal of Peace Research* 31, no. 4 (1994): 369–73, https://www.jstor.org/stable/424592; "Unsealed Soviet Archives Reveal Cover-Ups at Chernobyl Plant before Disaster," Reuters, April 26, 2021, https://www.reuters.com/world/unsealed-soviet-archives-reveal-cover-ups-chernobyl-plant-before-disaster-2021-04-26; Anna Hayes, "AIDS, Bloodheads & Cover-Ups: The 'Abc' of Henan's Aids Epidemic," *AQ: Australian Quarterly* 77, no. 3 (2005): 12–40, https://www.jstor.org/stable/20638337; Annie Sparrow, "The Chinese Government's Cover-Up Killed Health Care Workers Worldwide," *Foreign Policy*, February 6, 2024, https://foreignpolicy.com/2021/03/18/china-covid-19-killed-health-care-workers-worldwide; Ariana A. Berengaut, "Democracies Are Better at Fighting Outbreaks," *The Atlantic*, February 24, 2020, https://www.theatlantic.com/ideas/archive/2020/02/why-de-

mocracies-are-better-fighting-outbreaks/606976; Alastair Smith and Alejandro Quiroz Flores, "Disaster Politics: Why Natural Disasters Rock Democracies Less," *Foreign Affairs*, July 15, 2010, https://www.foreignaffairs.com/articles/2010-07-15/disaster-politics; and Michael Massing, "Does Democracy Avert Famine?" *The New York Times*, March 1, 2003, https://www.nytimes.com/2003/03/01/arts/does-democracy-avert-famine.html.

131. Vinand M. Nantulya and Michael R. Reich, "The Neglected Epidemic: Road Traffic Injuries in Developing Countries," *BMJ: British Medical Journal* 324, no. 7346 (May 11, 2002): 1139–41, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1123095.

132. Drexel and Kelley, "China Is Flirting with AI Catastrophe."

133. Sonia Ben Ouagrham-Gormley and Kathleen M. Vogel, "Follow the Money: What the Sources of Jiankui He's Funding Reveal about What Beijing Authorities Knew about Illegal CRISPR Babies, and When They Knew It," *Bulletin of the Atomic Scientists* 76, no. 4 (July 3, 2020): 192–99, https://doi.org/10.1080/00963402.2020.1780726; Kai-Fu Lee, *AI Superpowers: China, Silicon Valley, and the New World Order* (Boston: Houghton Mifflin Harcourt, 2018), 27, 102–3; and "State of AI Safety in China," Concordia AI, October 2023, https://concordia-ai.com.

134. "Global Opinions and Expectations about Artificial Intelligence," Ipsos, January 2022, https://www.ipsos.com/sites/default/files/ct/news/documents/2022-01/Global-opinions-and-expectations-about-AI-2022.pdf; "China's Grim History of Industrial Accidents," BBC News, December 21, 2015, https://www.bbc.com/news/world-asia-china-35149263; Clare Baldwin, Brenda Goh, and Sue-Lin Wong, "Corrected-RPT-China Chemical Safety Problems Highlighted before Tianjin Blasts," Reuters, August 24, 2015, https://www.reuters.com/article/idUSL3N10X057; Louise Moon, "Parents Haunted by China's Melamine Baby Milk Scandal Still Favour Foreign Brands," *South China Morning Post*, Business, February 22, 2020, https://www.scmp.com/business/companies/article/3051808/foreign-brands-still-dominate-parents-do-not-trust-chinas-home; and Sparrow, "The Chinese Government's Cover-Up Killed Health Care Workers Worldwide."

135. David Stanway, "Factbox: A History of China's Steel Sector," Reuters, May 3, 2012, https://www.reuters.com/article/idUSBRE84203A; Anatoly Zak, "Disaster at Xichang," *Smithsonian Magazine*, February 2013, https://www.smithsonianmag.com/air-space-magazine/disaster-at-xichang-2873673; and Ryan Dube and Gabriele Steinhauser, "China's Global Mega-Projects Are Falling Apart," *The Wall Street Journal*, January 20, 2023, https://www.wsj.com/articles/china-global-mega-projects-infrastructure-falling-apart-11674166180.

136. "U.S. Keeps Eye on China's Space Activities for Potential Risks," Associated Press, December 9, 2022, https://apnews.com/article/space-exploration-science-china-beijing-government-132254679381272785 2152bb4d65e08a; Zachary Cohen, "Exclusive: U.S. Assessing Reported Leak at Chinese Nuclear Power Facility," CNN, June 14, 2021, https://www.cnn.com/2021/06/14/politics/china-nuclear-reactor-leak-us-monitoring/index.html; David Stanway, "Explainer: What Happened at China's Taishan Nuclear Reactor?" Reuters, June 15, 2021, https://www.reuters.com/world/china/what-happened-chinas-taishan-nuclear-reactor-2021-06-15; Zeeya Merali, "Did China's Nuclear Tests Kill Thousands and Doom Future Generations?" *Scientific American*, July 1, 2009, https://www.scientificamerican.com/article/did-chinas-nuclear-tests; and Georgios Pappas, "The Lanzhou Brucella Leak: The Largest Laboratory Accident in the History of Infectious Diseases?" *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America* 75, no. 10 (November 14, 2022): 1845–47, https://doi.org/10.1093/cid/ciac463.

137. Leopold Aschenbrenner, "Nobody's on the Ball on AGI Alignment," For Our Posterity blog, March 29, 2023, https://www.forourposterity.com/nobodys-on-the-ball-on-agi-alignment; Katja Grace et al., "Thousands of AI Authors on the Future of AI" (arXiv, January 5, 2024), sec. 4.6, https://doi.org/10.48550/arXiv.2401.02843; and Amanda Askell, Miles Brundage, and Gillian Hadfield, "The Role of Cooperation in Responsible AI Development," arXiv, July 10, 2019, https://doi.org/10.48550/arXiv.1907.04534.

138. Jan Leike and Ilya Sutskever, "Introducing Superalignment," OpenAI blog, July 5, 2023, https://openai.com/blog/introducing-superalignment; "Frontier Risk and Preparedness"; Department for Science, Innovation and Technology, *Introducing the AI Safety Institute*, policy paper (London: UK Government, January 17, 2024), https://www.gov.uk/government/publications/ai-safety-institute-overview/introducing-the-ai-safety-institute; The White House, "Fact Sheet: Vice President Harris Announces New U.S. Initiatives to Advance the Safe and Responsible Use of Artificial Intelligence," press release, November 1, 2023, https://www.whitehouse.gov/briefing-room/statements-releases/2023/11/01/fact-sheet-vice-president-harris-announces-new-u-s-initiatives-to-advance-the-safe-and-responsible-use-of-artificial-intelligence; and Infocomm Media Development Authority, "First of Its Kind Generative AI Evaluation Sandbox for Trusted AI by AI Verify Foundation and IMDA," press release, October 31, 2023, https://www.imda.gov.sg/resources/press-releases-fact-sheets-and-speeches/press-releases/2023/generative-ai-evaluation-sandbox.

139. On adversarial exploitation of vulnerabilities, see: Vassilev et al., "Adversarial Machine Learning." For demonstrations of backdoored models creating cyber vulnerabilities, see: Evan Hubinger et al., "Sleeper Agents: Training Deceptive LLMs That Persist Through Safety Training," arXiv, January 11, 2024, https://doi.org/10.48550/arXiv.2401.05566.

140. Inez Miyamoto, "Disinformation: Policy Responses to Building Citizen Resiliency," *Connections* 20, no. 2 (2021): 52, https://doi.org/10.11610/Connections.20.2.05; "Countering Russian Disinformation," CSIS blog, September 23, 2020, https://www.csis.org/blogs/post-soviet-post/countering-russian-disinformation; and Rorry Daniels, "Taiwan's Unlikely Path to Public Trust Provides Lessons for the U.S.," The Brookings Institution, September 15, 2020, https://www.brookings.edu/articles/taiwans-unlikely-path-to-public-trust-provides-lessons-for-the-us.

141. "Democratic Roadmap: Building Civic Resilience to the Global Digital Information Manipulation Challenge," Bureau of Cyberspace and Digital Policy, U.S. Department of State, accessed March 29, 2024, https://www.state.gov/roadmap-info-integrity; Sander van der Linden, Stephan Lewandowsky, and Jon Roozenbeek, "Prebunking: How to Build Resilience against Online Misinformation," Context, September 5, 2022, https://www.context.news/digital-rights/opinion/prebunking-how-to-build-resilience-against-online-misinformation.

142. Daniel Armanios, Jonas Skovrup Christensen, and Andriy Tymoshenko, "What Ukraine Can Teach the World about Resilience and Civil Engineering," *Issues in Science and Technology* 40, no. 1 (Fall 2023): 98–103, https://doi.org/10.58875/URYE3161.

143. P. G. Sibly et al., "Structural Accidents and Their Causes," *Proceedings of the Institution of Civil Engineers* 62, no. 2 (May 1977): 191–208, https://doi.org/10.1680/iicep.1977.3222; Petroski, *To Forgive Design*, chap. 5.

144. Sean Brady, "The 30 Year Failure Cycle," Institution of Structural Engineers, May 1, 2013, https://www.bradyheywood.com.au/wp-content/uploads/2020/04/52fa7d_f1a1cdc27bf3431cb76e59eda74ae494.pdf; Petroski, *To Forgive Design*, chap. 9.

145. Perrow, *Normal Accidents*, chaps. 1–3.

146. Pranshu Verma and Will Oremus, "ChatGPT Invented a Sexual Harassment Scandal and Named a Real Law Prof as the Accused," *The Washington Post*, April 14, 2023, https://www.washingtonpost.com/technology/2023/04/05/chatgpt-lies.

147. Nitasha Tiku, Gerrit De Vynck, and Will Oremus, "Big Tech Was Moving Cautiously on AI. Then Came ChatGPT," *The Washington Post*, February 3, 2023, https://www.washingtonpost.com/technology/2023/01/27/chatgpt-google-meta.

148. Eoghan Stafford, Robert Trager, and Allen Dafoe, *Safety Not Guaranteed: International Strategic Dynamics of Risky Technology Races* (Oxford: Centre for the Governance of AI, November 1, 2022), https://www.governance.ai/research-paper/safety-not-guaranteed-international-strategic-dynamics-of-risky-technology-races.

149. Communications and Digital Committee, "Large Language Models and Generative AI," 37–39.

150. "We Need a Science of Evals," Apollo Research blog, January 22, 2024, https://www.apolloresearch.ai/blog/we-need-a-science-of-evals.

151. "Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy," Bureau of Arms Control, Deterrence, and Stability, U.S. Department of State, November 9, 2023, https://www.state.gov/political-declaration-on-responsible-military-use-of-artificial-intelligence-and-autonomy-2; Olivier Knox, "A.I. and Nuclear Decisions Shouldn't Mix, U.S. Says ahead of Biden-Xi Summit," *The Washington Post*, November 9, 2023, https://www.washingtonpost.com/politics/2023/11/09/ai-nuclear-decisions-shouldnt-mix-us-says-ahead-biden-xi-summit.

152. Drexel and Kelley, "Behind China's Plans to Build AI for the World."

153. Drexel and Kelley, "China Is Flirting with AI Catastrophe."

154. Cristina Criddle and Eleanor Olcott, "Chinese and Western Scientists Identify 'Red Lines' on AI Risks," *Financial Times*, March 18, 2024, https://www.ft.com/content/375f4e2d-1f72-49c8-b212-0ab2a173b8cb.

About the Center for a New American Security

The mission of the Center for a New American Security (CNAS) is to develop strong, pragmatic, and principled national security and defense policies. Building on the expertise and experience of its staff and advisors, CNAS engages policymakers, experts, and the public with innovative, fact-based research, ideas, and analysis to shape and elevate the national security debate. A key part of our mission is to inform and prepare the national security leaders of today and tomorrow.

CNAS is located in Washington, DC, and was established in February 2007 by cofounders Kurt M. Campbell and Michèle A. Flournoy. CNAS is a 501(c)3 tax-exempt nonprofit organization. Its research is independent and nonpartisan.

CNAS Editorial

DIRECTOR OF STUDIES
Paul Scharre

DEPUTY DIRECTOR OF STUDIES
Katherine L. Kuzminski

PUBLICATIONS & EDITORIAL DIRECTOR
Maura McCarthy

ASSOCIATE EDITOR
Caroline Steel

CREATIVE DIRECTOR
Melody Cook

DESIGNER
Rin Rothback

Cover Art & Production Notes

COVER ILLUSTRATION
Rin Rothback

COVER IMAGES
Christophe Simon/Getty, Sergei Supinsky/Getty, and Mass Communication Specialist 1st Class Ronald Gutridge/U.S. Navy

PRINTER
CSI Printing & Graphics
Printed on an HP Indigo Digital Press

Center for a New American Security
1701 Pennsylvania Ave NW
Suite 700
Washington, DC 20006
CNAS.org
@CNASdc

Contact Us
202.457.9400
info@cnas.org


Center for a
New American
Security