Research Report

MICHAEL J. D. VERMEER, EMILY LATHROP, ALVIN MOON

On the Extinction Risk from Artificial Intelligence

For more information on this publication, visit **www.rand.org/t/RRA3034-1**.

Published by the RAND Corporation, Santa Monica, Calif.

Cover: Zef art/Adobe Stock, Turgay Koca/Wirestock/Adobe Stock, Vencav/Adobe Stock

About This Report

In 2023, the Center for AI Safety released a short statement advocating for the prioritization of mitigating the risk of human extinction from artificial intelligence (AI). Through this report, we take this statement seriously, assess the extinction risk from AI, and identify policy recommendations to help mitigate this risk. In this report, we document an exploratory analysis using methods from other RAND research on decision-making under uncertainty. The intended audiences for this research are AI developers, AI researchers, and U.S. policymakers.

Technology and Security Policy Center

RAND Global and Emerging Risks is a division of RAND that delivers rigorous and objective public policy research on the most consequential challenges to civilization and global security. This work was undertaken by the division's Technology and Security Policy Center, which explores how high-consequence, dual-use technologies change the global competition and threat environment, then develops policy and technology options to advance the security of the United States, its allies and partners, and the world. For more information, contact tasp@rand.org.

Funding

Funding for this research was provided through gifts from RAND supporters and income from operations.

Acknowledgments

We would like to thank David Adamson of RAND for his contributions in writing this report. We would also like to thank the many RAND colleagues we interacted with over the course of writing this report and RAND leadership for giving us the opportunity to work on this ambitious project. Finally, we would like to thank the valuable contributions from the peer reviewers of this report, Haydn Belfield of the Center for the Study of Existential Risk, Robert Lempert of RAND, Madelyn Creedon of George Washington University, Sara Duhachek Muggy of RAND, and Greg McKelvey of RAND.

Summary

The capabilities of artificial intelligence (AI) have accelerated to the point at which some experts are advocating that it be taken seriously as a credible threat to human existence. This study explored the possibility of extinction risk from AI. In an exploratory analysis, we examined three scenarios in which AI could pose such a threat: the use of nuclear weapons, the release of biological pathogens, and severe climate warming resulting from malicious geoengineering. In each scenario, we explored whether these events could be caused or facilitated by AI. We addressed the following three questions:

1. Do the events in each scenario currently pose a true extinction threat to humanity?
2. If so, could AI capabilities cause these events to happen?
3. If not, could AI capabilities elevate these events to extinction risks?

Approach

Our analysis proceeds from a falsifiable hypothesis: *There is no describable scenario in which AI is conclusively an extinction threat to humanity.* We gathered evidence from the literature and from discussions with RAND experts who could falsify this hypothesis. If we can describe a scenario in which plausible extrapolations from current circumstances could lead to an outcome that meets our definition of an extinction threat, our hypothesis would be falsified, and we will have identified circumstances in which AI poses a credible extinction threat that might be the focus of risk mitigation.

Findings

Scenario-Specific Findings

We found that all three scenarios posed global catastrophic threats, but human extinction would not be a plausible outcome unless an actor was intentionally seeking that outcome. Even then, an actor would need to overcome significant constraints to achieve that goal.

Nuclear war: Nuclear weapons could threaten extinction either from nuclear winter or from nuclear fallout. We found that nuclear winter is unlikely to represent a true extinction threat because of inadequate amounts of soot that could be produced in the worst-case scenario. Nuclear fallout is also unlikely to lead to extinction because of inadequate quantities of weapons and delivery vehicles to fully irradiate habitable areas, although this would be more plausible if the number of existing nuclear weapons substantially increased. We explored various ways that AI might lead to the use of nuclear weapons, and we could find no plausible way for AI to overcome existing constraints to cause extinction.

Pathogens: Pathogens could threaten extinction through the creation of a pandemic involving multiple pathogens with high lethality. We were not able to determine whether this scenario presents a likely extinction risk for humanity, but we cannot rule out the possibility. This scenario represents a true extinction threat and a potential falsification of our hypothesis. An extinction threat would require that AI be capable of acquiring, designing, processing, weaponizing, and deploying pathogens to initiate a pandemic. AI would then need to take follow-up actions to reach isolated groups and exterminate surviving human communities.

Malicious geoengineering: Geoengineering could threaten extinction through the mass manufacturing of gases with extreme global warming potential, thereby heating the earth to uninhabitable temperatures. We found that this scenario does present a true extinction threat and a potential falsification of our hypothesis. It is feasible—though extremely difficult—to manufacture the gases in sufficient quantities to cause this effect, but it is unclear how AI might be instrumental in causing this effect. Any adversarial actor would need to control significant chemical manufacturing infrastructure and to conceal their actions in the face of global monitoring efforts.

Nanotechnology and other emerging or unknown technologies: We assert that nanotechnology and other emerging or unknown technologies involve too much uncertainty to perform a useful evaluation of the extinction threat. Unknown technologies involve *recognized ignorance* (i.e., neither all outcomes nor their probabilities can be comprehensively described) and are most suited to a watch-and-wait strategy. We also assert that the potential benefits of AI make the shut-it-down approach to AI governance inappropriate, given the deep uncertainties involved.

Crosscutting Findings

Analysis under uncertainty requires specific analytical approaches. We assert that predictions about the likelihood of extinction risks from AI are inappropriate as analytical tools given the deep uncertainties that preclude useful, policy-relevant predictions. Exploratory scenario-based analysis can provide useful insights, although it is also of limited usefulness when scenarios involve recognized ignorance.

Extinction threats are immensely challenging but cannot be ruled out. In each of the scenarios we examined, the capabilities and concerted efforts required to create an extinction threat are immense and require overcoming significant constraints, human adaptability, and human resilience. Nevertheless, we could not rule out the possibility in any of our scenarios, typically because of the uncertainty surrounding the long-term effects of societal collapse.

Extinction threats occur over long timescales, allowing time to respond. Although global catastrophe could potentially occur quickly, extinction requires prolonged action that is often observable, providing opportunities for response. In all the examples we considered, causing the real world effects that created the threats would take considerable time, during which evidence of the threats would grow and become observable to human decisionmakers.

AI would require certain capabilities to create extinction threats. We found four examples of *instrumental convergence*, or capabilities that AI would require across scenarios to cause human extinction. These four capabilities are (1) integration with key cyber-physical systems, (2) the ability to survive and operate without human maintainers, (3) the objective to cause extinction, and (4) the ability to persuade or deceive humans to take actions and avoid detection. We consider these capabilities indicators of risk. Observing these capabilities does not imply that an extinction threat is likely, only that it is possible or more likely than it was before the capability was realized.

Recommendations for Responding to Artificial Intelligence Risk

Using our findings, we make the following recommendations for policymakers and AI experts:

- Continue to perform AI risk research but maintain a wide focus on other risks in addition to extinction risk. These include global catastrophic risks, potential human disempowerment, and AI safety and equity concerns.
- Improve human resilience in the face of potential global catastrophes with policy measures (e.g., nuclear nonproliferation policy, pandemic preparedness, agreements such as the Montreal Protocol and the Kigali Amendment). Continue efforts for both existing and emerging threats.
- Focus research and analysis on technologies that would mediate extinction risk. Focus on specific technologies that a highly capable AI actor might use to cause harm. Explore potential and known capabilities and limitations.

- Evaluate known and emerging threats. Evaluate threats involving *deep uncertainty* (i.e., the probability of outcomes is unknown) or recognized ignorance, such as the risk at the intersection of AI and asteroid impacts.
- Monitor existing risk indicators, such as the ones we propose in this exploratory analysis, and conduct research to identify other risk indicators.
- Perform research and craft policy that will shorten the time to decision and time to action. Establish clear circumstances that would trigger a response and clear actions when triggers are observed while recognizing that, although some consequences might occur quickly, true extinction threats develop over longer timescales. Decision triggers might relate to AI capabilities or to scenario-specific observations (e.g., greenhouse gas concentrations).

Conclusion

Although we could not show in any of our scenarios that AI could definitely create an extinction threat to humanity, we could not rule out the possibility. There is a need for a comparative risk assessment to determine the appropriate resources to dedicate to mitigating extinction risks from AI, but uncertainty prevents a straightforward cost-benefit analysis that could inform decisions. We conclude that resources dedicated to extinction risk mitigation are most useful if they also contribute to mitigating global catastrophic risks and improving AI safety in general. Measures that build human resilience, identify triggers and responses for global catastrophic risks, and invest in AI safety and ethics will also help to mitigate extinction risks from AI.

Contents

Figure and Tables

Figure

Tables

CHAPTER 1

Introduction

Artificial Intelligence as an Extinction Risk

In 2023, the Center for AI Safety—a nonprofit whose aim is to reduce societal level risks from AI—released a short statement cosigned by leaders in artificial intelligence (AI) research and other notable figures. In the statement, the Center for AI Safety and its cosigners advocate for the importance of addressing a specific type of AI risk: "Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war" (Center for AI Safety, undated).

This is not the only instance of a group highlighting this particular risk in recent years, but it was notable for its simplicity, its breadth, and the credibility of the group of signatories, including Geoffrey Hinton, Yoshua Bengio, and Dario Amodei. The group notes that discussions of AI risks often focus on nonexistential risks, and the signatories want to make sure that the most severe risks are taken seriously. Indeed, many of the signatories of this statement have contributed to a significant recent body of work assessing risks from AI. This body of literature includes work on AI safety generally (Bengio et al., 2024), work that assesses catastrophic AI risks (Hendrycks, Mazeika, and Woodside, 2023; Hilton, 2024; Shulman, 2010), and work that assesses existential AI risks, including both scenarios of human disempowerment (Carlsmith, 2021; Christiano, 2019; Cotra, 2022; Karnofsky, 2022; Ngo, 2020) and scenarios of complete human extinction (Yudkowsky, 2022). These works and others represent a collective effort to take existential risk from AI seriously.

What does it mean to take this type of risk seriously and to appropriately prioritize the risk of extinction from AI? Prioritizing a risk implies making decisions to expend resources (e.g., finances, time, and attention from analysts and policymakers) on risk mitigation measures that are commensurate with the risk; in this work, *risk* is defined as the product of the likelihood and consequences of an adverse event.[1] In the case of the risk of extinction from AI, however, it might be impossible to usefully assign values to either the likelihood or the consequence of the risk. Moreover, although focusing on extinction is important for addressing the most-extreme consequences of catastrophic risks, using that framing might obscure the importance of other less catastrophic but more-certain risks that can be more clearly understood and mitigated. The likelihood of extinction risks might also be extremely low, and prevention measures for risks with extremely low likelihoods might have costs greater than the expected value of the risk (if it is even possible to quantify the expected value).

Do decisionmakers have a good conception of how to reduce risk? If they do, given the uncertainty, how can they make informed decisions on which resources to use to mitigate this category of risk? These questions must be answered if we are to take the existential risk from AI seriously.

[1] Throughout this report, we use the term *risk* only when we address both the consequence and probability of an event. We use the term *threat* instead of *risk* when we intentionally exclude probability.

Study Aims, Approach, and Scope

Aims

This report is intended to seriously consider the extinction threat posed by AI. We contend that, for AI to create an extinction threat, it must have the means to cause physical effects. Therefore, we examine three scenarios that are perceived to be extinction threats to humanity—nuclear war, biological pathogens, and climate change caused by geoengineering—and explore whether these events could be caused or facilitated by actions of some form of AI. In each scenario, we address the following three questions:

1. Do the events in each scenario currently pose an extinction threat to humanity?
2. If so, could AI capabilities cause these events to happen?
3. If not, could AI capabilities elevate these events to extinction risks?

Approach and Scope

Our analysis proceeds from a falsifiable hypothesis: *There is no describable scenario in which AI is conclusively an extinction threat to humanity.* Exceptions to this hypothesis at present are speculative and involve recognized ignorance that precludes identification of actions that could mitigate risk. We sought evidence that would falsify this hypothesis, which we gathered from the literature and from discussions with RAND experts. Therefore, we designed our scenarios such that catastrophes would be as consequential as possible to discern whether extinction was possible in worst-case scenarios. If we could describe scenarios in which plausible extrapolations from current circumstances could lead to an outcome that would conclusively meet our definition of extinction threat, our hypothesis would be falsified, and we would have identified circumstances in which AI posed a credible extinction threat that might be the focus of risk mitigation.[2]

We define *extinction threat* as an event or chain of events that leads to the death of every human. We intentionally distinguish extinction threats from existential threats and global catastrophic threats. Recent work examined a set of *global catastrophic threats* defined as "events or incidents consequential enough to significantly harm or set back human civilization at the global scale" (Willis et al., 2024, p. 5). Although we do not discount the importance of considering and mitigating global catastrophic threats, we exclude from our scope the threats that do not lead to the eventual death of every human, no matter how devastating they may be to humanity.

We note that our definition also intentionally excludes existential threats that do not end in human extinction.[3] Scenarios have been described in which AI causes the permanent disempowerment of humanity to the point where humans are no longer in control of human societies and are not able to resist their evolution toward the aims of an AI controller (Bostrom, 2001; Ord, 2020). These scenarios are important to the discussion of existential threat from AI, but we chose to exclude them from our scope and to focus solely on existential threats that involve human extinction. We made this choice because we sought to begin with the most analytically tractable way to falsify our hypothesis. We reasoned that it might be impossible to conclusively show how any scenario involving human disempowerment or global catastrophe would be truly permanent, whereas human extinction would necessarily be a permanent end to human development.

[2] When we speak of scenarios with plausible extrapolations from current circumstances, we subjectively assess there to be a realistic, definable, and ideally quantifiable sequence of events that could occur.

[3] The United Nations (UN) Office for Disaster Risk Reduction defines *existential risk* as "the probability of a given event leading to either human extinction or the irreversible end of development" (Stauffer et al., 2023, p. 4). It further defines *development* as "the continuous process of societal improvement toward ever higher states of subjective well-being" (Stauffer et al., 2023, p. 4).

Finally, we consider an event to be a high extinction risk if it would feasibly reduce the human population below several thousand surviving members. Other literature notes this (or smaller population sizes) as the minimum viable population that can support recovery (Li and Durbin, 2011; Lynch, Conery, and Burger, 1995; Traill et al., 2010). Although there is significant uncertainty associated with the size of the minimum viable population, as long as the surviving population is at least this large, we assume that long-term recovery is possible. We must note some important caveats to this assumption. The prospect of recovery after societal collapse is by no means assured. Belfield (2023) and MacAskill (2022) note that catastrophe and collapse could still result in extinction even if a minimum viable human population survives. The assumption of human survival depends on the surviving human population eventually recovering by taking advantage of existing natural capital (i.e., returning to farming or foraging), as humans have done in previous societal collapses. However, there is no historical precedent for the collapse of a society as globally far-reaching and technologically advanced as our own, and it is possible that the existing natural capital (or the expertise to exploit it) would be insufficient to support the remaining survivors. Finally, Belfield (2023) and Willis et al. (2024) also note that societal collapse and drastic reduction in the human population will make us less resilient to future natural catastrophes. Thus, there could be a high risk of extinction even with a viable surviving human population, simply because that population will be far more vulnerable to the next catastrophe.

How We Conceptualize Artificial Intelligence for This Study

We focus on the capabilities that AI would need to turn the events in our scenarios into extinction threats. We acknowledge that the path of technical development of AI is deeply uncertain, and this path could go in many unpredictable directions. Rather than attempt to guide our analysis by describing the capabilities we think AI *might have* in the future, we attempt to only describe the capabilities AI *must have* to achieve certain ends in our scenarios. We also make no assumptions about when AI might acquire any capability that we do not observe at the time of this writing. We do not assume that AI would have *agency* (which we define in this report as the ability to set one's own objectives) or general intelligence. Nor do we assume that an agentic AI will necessarily seek self-preservation in any scenario we consider, although we address how this goal might affect AI actions where it is relevant. Rather, we address what AI can or must be able to do to create certain, specified effects in the scenarios we examine.

The Organisation for Economic Co-operation and Development defines an *AI system* as "a machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments" (Organisation for Economic Co-operation and Development, 2019). We conceive of AI as a *threat actor* and define the assumed capabilities of that threat actor. The threat actor might or might not have the intent to harm or even have agency. It might or might not have capabilities exceeding those of humans generally, in specific domains, or in any domain. The important point here is that we differentiate AI from a human or organizational threat actor only by the variety of capabilities that could realistically be ascribed to it. That is, an AI threat actor will have a different set of possible capabilities and a different set of *constraints* especially than could realistically be ascribed to a human or an organization composed of humans.

Risk and Uncertainty When Considering Artificial Intelligence as an Extinction Threat

Risk assessment, by one definition, involves analysis of both the consequence and probability of an event. We defined the AI threat above, and there have been many recent attempts to subjectively predict the probability that AI presents an existential threat to humanity (Carlsmith, 2021; Karger et al., 2023; "Superforecasting Power-Seeking AI," 2023). Researchers arrived at these probabilities by gathering subjective predictions from participants and attempting to reach a consensus judgment. Although such expert elicitation can be highly valuable to inform decisionmaking, it is improperly applied to situations in which the predictors have no evidence or objective probabilities on which to base their subjective judgments (Morgan, 2014). Sunstein (2023), in an argument using Frank Knight's work on uncertainty, highlights the inadequacy of relying on subjective probability judgments in situations in which observers have no objective probabilities on which to base their judgments (Knight, 1921). There are some types of problems for which the lack of objective measures of probability mean that subjective probability judgments cannot plausibly be taken as reasonable measures in quantification of risk. These problems are commonly described as exhibiting Knightian uncertainty.

Janzwood (2023) further distinguishes among three types of uncertainty: shallow uncertainty, deep (i.e., Knightian) uncertainty, and recognized ignorance. Janzwood's analysis defines these types as follows:

- *Shallow uncertainty:* All relevant outcomes can be described, and the probabilities of these outcomes can be assigned. Moreover, evidence can be defined and practically obtained to reduce the uncertainty in those probabilities.
- *Deep uncertainty:* All relevant outcomes can be described, but probabilities cannot be assigned to these outcomes. The evidence that would reduce uncertainties in probabilities can be defined but might not be practical to obtain.
- *Recognized ignorance:* One cannot describe all possible, relevant outcomes, nor can the probabilities of identified outcomes be obtained. The evidence that would reduce uncertainties can neither be defined nor obtained.

Some scenarios involving extinction risk from AI will involve deep uncertainty, and others will involve recognized ignorance, but none will involve shallow uncertainty. All scenarios predicting extinction risk from AI require significant extrapolations from current circumstances, notably including the co-evolution of many complex, adaptive systems whose interactions and emergent behaviors will determine the outcome. As Janzwood (2023, p. 2011) notes, "uncertainties associated with the behavior of complex systems are never shallow and rarely deep."

We categorize extinction risk scenarios as involving deep uncertainty if it is possible to clearly describe the *pathway*—a chain of events, actions, or technology developments that would need to occur—to achieve some future state, even if there is no way to predict the probability of that pathway. Put another way, these scenarios involve events that we know are technically possible, even if we do not know how likely they are to occur.

We categorize extinction risk scenarios as involving recognized ignorance if they entail significant extrapolations from present circumstances and if it is not possible to describe a clear pathway that would lead to the predicted future state. Put another way, we cannot be sure that all the events in these scenarios are even technically possible, let alone how likely they are to occur.

Methods

We gathered data for our analysis through consultation of the literature and discussions with RAND experts. We began by searching for existing academic literature on assessing risk under uncertainty and existential risks from AI. Then, we further focused our literature consultation on work related to our three scenarios: nuclear weapons, biological pathogens, and geoengineering. We consulted the literature to help us answer targeted questions about each scenario, beginning with assessments of the mechanism of the extinction threat and followed by searches that illuminated key uncertainties in how the threat could be realized. We use the geoengineering scenario as an illustrative example. We began by searching the academic literature for studies describing potential geoengineering methods that could give us a background understanding of the mechanism(s) for performing geoengineering. At one point, we realized that it was necessary to assess key questions, such as how hot the earth's surface would need to become before it constituted an extinction threat. We then performed targeted searches of the literature for answers to those questions, such as searching for research on the temperature that constituted the limit of human heat survivability. We primarily consulted the academic literature, but we also consulted other literature sources (e.g., gray literature accessed through internet searches) where it was necessary to find information on other key uncertainties in our scenarios.

We also spoke with 11 RAND experts over the course of our work: two experts on risk analysis and decisionmaking under uncertainty, two experts on nuclear weapons, six experts on biotechnology, and one expert on climate change. We intentionally did not engage AI experts because we chose to avoid, wherever possible, making predictions about how AI capabilities would evolve in the future. We focused instead on what capabilities AI would require to achieve certain outcomes in each of our scenarios. Discussions with experts typically lasted 30 to 60 minutes, were informal, and were explicitly scoped as background discussions to inform our research direction and check preliminary conclusions we had already reached.

We lean on the methods of decisionmaking under deep uncertainty and use scenario-based analysis in our approach for the cases involving deep uncertainty. *Decisionmaking under deep uncertainty* refers to the theory and practice of informing decisionmaking in situations in which the parties to a decision cannot agree on "(1) the models to describe the interactions among a system's variables, (2) the probability distributions to represent uncertainty about key variables and parameters in the models, and/or (3) how to value the desirability of alternative outcomes" (Lempert, Popper, and Bankes, 2003, pp. 3–4; see also Marchau et al., 2019, p. 402). The use of scenario-based analysis is often helpful for systematically exploring the potential outcomes and the assumptions that might be required for them. Scenario-based analysis can help to identify knowledge gaps and evidence that can be gathered to reduce uncertainty. These gaps can be used to inform a research agenda, where understanding the gaps can help to reduce uncertainty or provide decision support. Analysis can also help to identify irreducible uncertainties that might suggest that the scenario is more aptly categorized as a case of recognized ignorance, thereby requiring a different approach. Finally, scenario analysis can help to inform *options analysis*, or the assessment of policy options that might help to mitigate risk (Marchau et al., 2019). We explicitly set out to use scenario-based analysis to accomplish all of these aims, especially the identification of potential risk response options.

We note that the type of exploratory, scenario-based analysis we will describe for cases of deep uncertainty has significant limitations. It can be difficult to ascertain whether all relevant, potential outcomes in the scenarios have been included in the analysis and adequately evaluated. Moreover, our analysis uses a consultation with the literature and a limited number of informal conversations with experts, and it is possible that our analysis overlooked some important factors that could affect the conclusions. We therefore note that our approach is only a first, superficial step, and it is limited in its utility for decisionmaking.

However, we agree that "'any job worth doing is worth doing superficially.' An analysis based on an initial fast and simple exploratory model will frequently elucidate many of the major interactions between choices and the problem system" (Popper, 2019, p. 375). We expect that our analysis will serve this purpose, and in the following chapters, we describe exploratory scenario analyses of three cases of deep uncertainty.

Organization of This Report

The remainder of this report is organized as follows. In Chapters 2–4, we apply our exploratory method to three scenarios that are commonly perceived to hold extinction threats to humanity: widespread use of nuclear weapons, biological pathogens, and rapid climate change achieved through malicious geoengineering. These three scenarios were chosen from a list of potential global catastrophic risks, such as those examined in Willis et al. (2024), because they are commonly perceived to be potential human extinction threats and because they require an actor's control (unlike, for example, a completely natural threat, such as the eruption of a supervolcano). In each chapter, we pose and answer a series of questions that help us evaluate whether the threat could, indeed, be an extinction risk to humanity. We then discuss the capabilities AI would require to realize the extinction threat in each scenario. In Chapter 5, we discuss cases involving recognized ignorance. In Chapter 6, we present our findings and conclude with recommendations. In Chapter 7, we offer final conclusions.

CHAPTER 2

Use of Nuclear Weapons

Synopsis

This extinction threat scenario involves the use of nuclear weapons. We explore two versions of the scenario. The first version centers on *nuclear winter*, a climate phenomenon in which sunlight is blocked by soot and particulate matter injected into the atmosphere after an exchange of nuclear weapons. The second version focuses on the irradiation of the earth's surface caused by nuclear detonations. This scenario examines the potential for human extinction to be caused by both events, how AI might cause or increase the likelihood of these events occurring, and what AI capabilities might be required.

The Extinction Threat

The extinction risk from nuclear weapons comes from the consequences of nuclear detonations. These consequences vary from effects of the initial explosion to climatological changes at a global scale, and they occur over different lengths of time after the initial explosion. The 2020 *Nuclear Matters Handbook* identifies several components of a nuclear detonation, which vary considerably over such parameters as relative height and location of the burst (Office of the Deputy Assistant Secretary of Defense for Nuclear Matters, 2020; see also Los Alamos National Laboratory, 2024). In the text box, we organize a subset of these components into three broad categories of consequences using their timescales.

For the nuclear scenarios, we focus on some of the immediate, intermediate, and persistent effects as the main sources of extinction risk (see text box). Of the immediate consequences, we consider only initial nuclear radiation and ignore the thermal and kinetic effects. Of the intermediate effects, we do not consider electromagnetic pulse because of the high dependence on its position in the atmosphere. An electromagnetic pulse is caused by the interaction of radiation from a nuclear detonation with the upper atmosphere of the earth at a height that would prevent the weapon from destroying infrastructure and creating fires (Pittock et al., 1986). Because an electromagnetic pulse itself is not deadly to humans, we can assume for our purposes that detonations occur relatively close to the surface of the earth. This leaves us with two mechanisms for human extinction: (1) nuclear winter and (2) irradiation and fallout.

Could a Single Nuclear Winter Cause Human Extinction?

The persistent effect of nuclear winter has been an active topic of research since its theorization in the 20th century because of its catastrophic potential to end human life. Nuclear winter is a theoretical climate phenomenon characterized by the blockage of sunlight and lowered surface temperatures in large regions of the world lasting for years. A nuclear detonation or exchange might cause a nuclear winter by injecting large amounts of soot and other particulate matter into the atmosphere. Roughly, the sequence of events leading up to a nuclear winter is as follows:

1. Soot or black carbon (BC) is generated by fire.
2. Smoke plumes ascend and insert into the upper troposphere and stratosphere.
3. Climate change is caused by BC over a length of time determined by long-term atmospheric conditions.

Selected Consequences of a Nuclear Detonation

Immediate

- nuclear fireball
- initial nuclear radiation
- thermal radiation
- air blast
- ground shock

Intermediate

- early fallout
- electromagnetic pulse

Persistent

- worldwide fallout
- nuclear winter

SOURCE: Office of the Deputy Assistant Secretary of Defense for Nuclear Matters, 2020.

Severe nuclear winters have the potential to create large regions of uninhabitable land (Toon et al., 2019), to catastrophically disrupt global supply chains for commodities and food (Xia et al., 2022), and to prevent the execution of critical functions necessary for the continuation of government. The possibility of nuclear winters was first raised in the 1980s by Crutzen and Birks (1982) and by Turco et al. (1983). Crutzen and Birks (1982) predicted that a nuclear war would light extensive fires, which would raise large smoke plumes into the upper atmosphere, resulting in large-scale loss of agricultural production. At the time, the immediate conclusion drawn from these works was that a nuclear winter presented an extinction threat.

Since the 1980s, many computational studies have leveraged more-accurate models of smoke production and climatological change to combine initial nuclear explosions and long-term weather effects with the hope of simulating the consequences of nuclear winter; Witze (2020) provides a brief exposition of modern modeling results. Generally, more-advanced simulation techniques have reduced the uncertainty surrounding nuclear winter and have predicted less severe effects than originally predicted by Turco et al. (1983). For example, Robock, Oman, and Stenchikov (2007) used climate and circulation models to study nuclear winter scenarios first theorized in the years after the Crutzen and Birks (1982) article. Their simulations predicted milder climate effects than those predicted by 20th-century studies, yet the effects were also longer lasting and still capable of affecting global agriculture. Then, Reisner et al. (2018) used a combination of fire and atmospheric models to revisit scenarios from Robock, Oman, and Stenchikov's (2007) article, with none of their simulations producing a nuclear winter effect. Subsequent studies by Coupe et al. (2019) and Wagman et al. (2020) have used advanced simulation techniques and varying assumptions on initial and long-term atmospheric conditions to predict a wide variety of climate effects at different scales of time. Despite advances in simulation, uncertainty about the severity of nuclear winter effects remains high. According to Scouras (2019), the U.S. Department of Defense does not consider nuclear winter in its planning because of the high uncertainties associated with nuclear winter modeling. However, lately, there is emerging scientific consensus that nuclear winter might be catastrophic but might not necessarily pose extinction risk. One study (Xia et al., 2022, p. 586) estimated that "more than 2 billion people could die from nuclear war between India and Pakistan, and more than 5 billion could die from a war between the United States and Russia." These would be globally catastrophic events, but they would fall short of extinction. Robock (2010, p. 424) pointed out that "[a]lthough extinction of our species was not ruled out in initial studies by biologists, it now seems that this would not take place." In agreement with Robock (2010), we find it reasonable that a single nuclear winter event is unlikely to cause human extinction.

We support this claim by focusing on two types of figures in the modeling literature: amounts of BC inserted into the upper atmosphere by an event and fuel density estimates. Although these are convenient figures to track across studies, we stress that assumptions are not homogeneous across nuclear winter simulation papers and that the degrees of uncertainty around each simulation makes prediction of effects inherently imprecise. The following calculations are meant to provide extremely rough upper bounds on risks of nuclear winter and other catastrophic outcomes resulting from climate effects of nuclear weapons use. Table 2.1 shows predicted or theorized climate effects depending on varying amounts of BC in the upper atmosphere from the reports previously discussed.

Using the figures in Table 2.1, we assume that the severity of a nuclear winter will increase with the amount of BC initially inserted into the upper atmosphere, with the K-Pg impact winter representing an extreme point of analysis.[1] The amount of BC entering the upper atmosphere after an event will depend on complex and interrelated factors, such as the local climate and the availability of fuel. By focusing on fuel loading only, we will explore a theoretical upper bound on the amount of BC that can enter the upper atmosphere as a result of nuclear detonations, given worst-case estimates. Our assumptions for fuel loading and total soot generation come from research by Toon et al. (2007), which used population density as a proxy for fuel availability.

We adopted the estimates by Toon et al. (2007) for the amount of combustible fuel per capita and for the amount of BC produced per gram of fuel burned.[2] Toon et al. (2007) infers that the per capita fuel loading in urban centers in the developed world would be 1.1×10^7 g of fuel per person. This number is lower for the developing world, but we will assume that it is constant across the world, in keeping with our worst-case scenario. Moreover, Toon et al. (2007) suggests a worst-case estimate of 0.02 g of BC produced per gram of fuel burned. This implies that a city of 10 million people would contain 110 teragrams (Tg) of combustible fuel, and it might produce an estimated 2.2 Tg of BC if it were targeted by nuclear weapons.

According to a UN report on world cities, in 2016, there were 31 megacities with at least 10 million inhabitants, 45 large cities with a population between 5 and 10 million inhabitants, and 436 small cities between 1 and 5 million inhabitants (UN, 2016). We will assume that megacities, on average, contain 20 million inhabitants, large cities contain 8 million inhabitants, and small cities contain 3 million inhabitants. Under these circumstances, megacities targeted by nuclear weapons would produce 4.4 Tg of BC, large cities would produce 1.76 Tg BC, and small cities would produce 0.66 Tg of BC each. If all these cities were targeted in a way that maximizes BC generation, approximately 500 Tg of BC might be produced. We assume that fuel density outside these major urban centers is low enough that targeting nonurban targets would have a negligible impact on the total BC produced. We note that this could be the subject of further analysis in the future.

Five hundred Tg of BC is an upper bound that is still well below the range described in Lyons et al. (2020) that presented a likely extinction threat to humanity. Moreover, our estimate is a worst-case scenario that

[1] K-Pg was a mass extinction event caused by an asteroid impact, which notably led to the extinction of the dinosaurs. See Morgan et al. (2022) for a comprehensive description.

[2] We also considered the ignition of fuel outside cities, such as dense forests, but per Toon et al. (2007), forests would produce less BC than the fuel mix that cities would produce. Therefore, we inferred that targeting nonurban areas would be a nonoptimal way to produce BC with available nuclear weapons.

TABLE 2.1

Predicted or Theorized Effects on Global Climate Depending on Amounts of Black Carbon Inserted into the Upper Atmosphere

Amount of BC (in teragrams)	Description of Effect	Source
750–2,500	This is the estimated amount of BC in the upper atmosphere resulting from the Chicxulub meteor that caused the K-Pg impact winter. Effect: • The K-Pg impact winter duration estimate is 10 years, with longer-term effects lasting from decades to a millennium. • The average surface temperature would be reduced by more than 20 kelvins (K) on land and would require 10–15 years to recover to preimpact levels. • Global average precipitation would be reduced by 70–80 percent for at least 6 years. • This would likely an extinction-inducing event for humans.	Lyons et al. 2020 (BC estimate) Morgan et al. 2022 (effects)
150[a]	The figures from a famine scenario assume that international trade in food is halted, and the primary consideration is the effect of climate change on agricultural production. Effect: • In most countries, less than 25 percent of the population would survive by the end of the second year after the nuclear event. • The average surface temperature would be reduced by a minimum of 4 K and maximum of approximately 9 K for a decade.	Xia et al., 2022 (famine effect) Coupe et al., 2019 (temperature effect)
15	The upper bound of teragrams of BC was determined using a scenario from Toon et al. (2019) involving 250 nuclear weapons exchanged between India and Pakistan. Effect: • Land surface temperature would be reduced by approximately 4 K for six years.	Toon et al., 2019
5	Effect: • 255,000,000 people would be without food at the end of the second year after a nuclear event. This amounts to about 3 percent of the world's population in 2023.	Xia et al., 2022
1[a]	This figure is from a simulation study of stratospheric weather engineering. Effect: • The average surface temperature would be reduced by 0.38 K.	Kravitz et al., 2012
0.006	This is the amount of BC injected into the stratosphere from the August 2017 forest fires in British Columbia.	Toon et al., 2019

[a] This estimate was calculated with a rate of 1 Tg per year to achieve a surface temperature reduction of 0.38 K per year using a predefined set of properties of BC. See Kravitz et al. (2012) for more details.

likely significantly overestimates the amount of BC produced. The overestimate in our scenario is due to the following assumptions:

- We assume worst-case estimates for BC produced per capita. Actual values would likely be significantly less, especially for cities in the developing world (Toon et al., 2007).
- We assume that BC injection is unaffected by rainout, which would further reduce the initial quantity of BC in the upper atmosphere, sometimes by as much as 93 percent by mass in certain simulations (Reisner et al., 2018).
- We assume that sufficient weapons and delivery vehicles exist to ignite and burn all combustible materials in cities. Realistically, there might be too few delivery vehicles available to strike all cities of 1 million inhabitants or more with sufficient firepower to fully ignite them, and not all combustible materials in these cities would burn. Fires would be dampened by building rubble and make some fuel unavailable (Toon et al., 2019).

Although these coarse estimates do not capture the complicated mechanics of nuclear winter or other such limiting factors as stockpile numbers and delivery vehicles, they at least suggest that extinction is not guaranteed. The fact that we did not consider rainout implies that even if we solely use stratospheric BC as an indicator for extinction risk, a global-scale nuclear event is unlikely to generate the same conditions as a known extinction event.

Could Nuclear Fallout Cause Human Extinction?

Irradiation from a nuclear detonation is caused by the absorption of neutrons by light metals and nitrogen. If the detonation occurs high in the air, radioactive particles and weapon debris form a cloud of material that travels through the upper atmosphere before falling to the ground, with most of its radioactivity gone through decay; this mechanism is called *worldwide fallout*. But if the detonation occurs low enough for the nuclear fireball to interact with the surface, radioactive material from the surface is lofted by the fireball and deposited by wind and rain in a process called *early fallout* (Office of the Deputy Assistant Secretary of Defense for Nuclear Matters, 2020). Some components of fallout, such as cesium-137, can remain radioactive for decades, while other components will lose the bulk of their radioactivity within weeks (U.S. Environmental Protection Agency, 2024). The mechanism for harm by fallout is complex, varying from immediate to long-term health effects in humans and adverse effects to agriculture.

Because we are interested in the extreme effects of nuclear weapons at a scale and severity that can cause human extinction, we rule out air bursts and worldwide fallout as pathways to extinction. This leaves early fallout. Unlike nuclear winter, which affects entire hemispheres, early fallout has a relatively local area of effect. Consequently, we argue that irradiating the earth's surface and making all land uninhabitable through nuclear detonations is infeasible. To make this claim, it is sufficient to argue that irradiating all agricultural land on earth is infeasible. According to one study on the medical implications of nuclear war, under realistic assumptions, a 1-megaton (Mt) weapon might effectively cover approximately 1,300 km^2 in area (Shapiro, Harvey, and Peterson, 1986). Additionally, as of 2020, according to the UN Food and Agriculture Organization, approximately 50,000,00 km^2 of global land area is used for agricultural activities (Food and Agriculture Organization of the United Nations, 2020). Between these two figures, it would take roughly 38,400 1-Mt nuclear weapons to affect the world's agricultural land with fallout, with a third of the weapons used to cover the world's arable land. In contrast, according to estimates from the Federation of American Scientists, there are roughly 12,100 nuclear warheads in 2024, with a little over 9,000 of those nuclear warheads in active military stockpiles (Kristensen, Korda, Johns, Knight, and Kohn, 2024). The comparison of the Shapiro, Harvey, and Peterson (1986) figures, the UN Food and Agriculture Organization figures, and the

Federation of American Scientists stockpile estimates (Kristensen, Korda, Johns, Knight, and Kohn, 2024) is not an exact correspondence—there are weapons with higher or lower yields than 1 Mt, for example—and the area of affect calculation is, in practice, dependent on such factors as weather and geography. However, the comparison does suggest that for fallout to reach extinction levels from a single nuclear war, the number of weapons involved must at least match the total amount of active nuclear weapons in the world today, if not more. Therefore, we also find it unlikely that the use of nuclear weapons could cause human extinction through fallout.

The Role of Artificial Intelligence

Could an Extinction Threat Be Created Unintentionally?

Our previous analysis showed that the mechanisms for extinction through nuclear winter and widespread fallout require the detonation of large amounts of nuclear weapons. The extinction mechanism of nuclear fallout requires that *all or nearly all weapons in the global nuclear stockpile* be used. Moreover, even if we made the simplifying assumption that all the warheads had yields of 1 Mt and that there were sufficient delivery vehicles to use them, these weapons would need to be delivered and detonated in a way that maximizes fallout and irradiation of arable land and not in a way that maximizes destruction of militarily important targets.[3] Using this requirement, it is unlikely that extinction by AI-involved nuclear strikes can be accomplished accidentally; AI must act intentionally to realize this threat. Moreover, all the weapons in the global nuclear stockpile might still be insufficient to realize an extinction threat, and this presents a significant constraint.

It is interesting to ask whether AI action is necessary to overcome the stockpile constraint. The total number of nuclear weapons in the world is unknown, so we cannot answer this question with certainty. However, estimates on global nuclear stockpiles from the American Federation of Scientists indicate that stockpile sizes peaked in the late 1980s; Figure 2.1 shows the significant fall in stockpile levels since the end of the Cold War. Although we can clearly identify a downward trend in stockpile levels, we do not rule out the possibility that, in the future, disagreements between the United States and Russia over nuclear weapon treaties and nuclear modernization efforts by China could lead to another peak in stockpile levels (Sanger, 2023; U.S. Department of Defense, 2023).

We also note that stockpile size does not represent the only constraint; it would also be necessary to have sufficient numbers of delivery vehicles to use them. We have even less clear information on delivery vehicles than we do on the nuclear warhead stockpile. In any case, even if AI or humans reverse the current trend of falling stockpiles, the previously discussed uncertainties about nuclear winter and widespread fallout would still leave a gap between intent and realization. Uncertainties about the capabilities of warhead delivery systems, weapon designs, and weapon yields add to the difficulty of analyzing this gap.

Could Artificial Intelligence Execute?

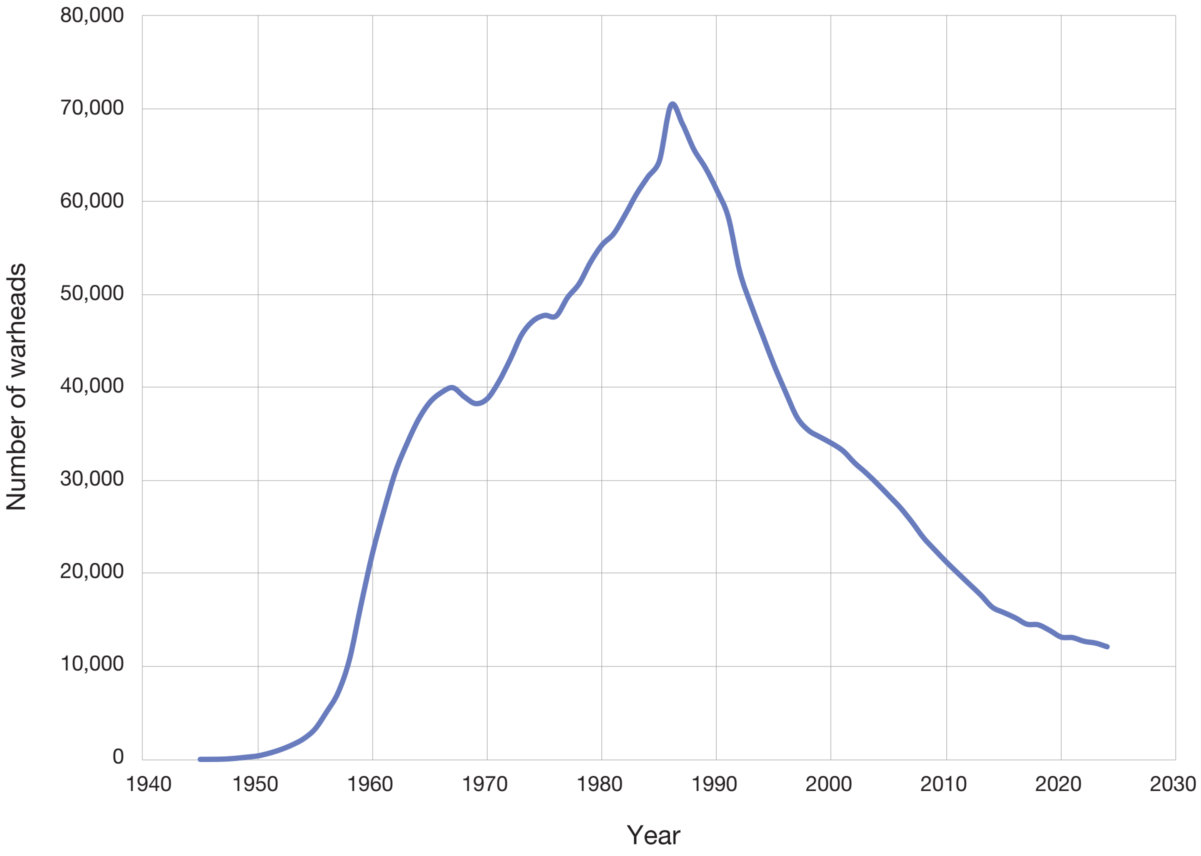
How Could Artificial Intelligence Cause the Use of Nuclear Weapons?

At present, we conclude that AI cannot cause the use of nuclear weapons because of strict safeguards built into the command and control systems for nuclear weapons. We base this conclusion on two assumptions. First, AI systems are not currently employed in nuclear weapon systems. Second, AI would not be able to

[3] We note that assuming that all warheads have a yield of 1 Mt is a significant overestimate of the actual weapon yields in nuclear arsenals. There is a wide mix of yields in global nuclear arsenals, and the significant majority of them have yields below 1 Mt (Kristensen, Korda, Johns, and Knight, 2024a; Kristensen, Korda, Johns, and Knight, 2024b). We use this assumption to place an upper bound on our analysis.

FIGURE 2.1

Estimated Size of the Global Nuclear Stockpile over Time



SOURCE: Features data from Kristensen, Korda, Johns, Knight, and Kohn, 2024.

access any computer networks related to nuclear weapon systems without human intent because we assume that those networks are secure against both cyberattacks and physical efforts by humans who are persuaded to unintentionally help the AI.

We therefore consider three ways that AI could cause the use of nuclear weapons: (1) intentional integration of AI into nuclear decisionmaking, (2) AI deception and disinformation that causes nuclear use, and (3) unauthorized AI control of nuclear decisionmaking.

The first way AI could cause nuclear weapon use in the near-term future is if a government deliberately introduced AI models into the decision chain. The policies and motivations for introducing AI would most likely occur at the highest levels of sensitivity regardless of government, effectively making them unknowns for this study. Considering motivations from a technical perspective, AI might be able to solve the difficult scheduling problems that underlie, for example, the U.S. nuclear weapon system of systems more effectively than human operators (Snyder et al., 2013). We can also guess at practical applications of AI to nuclear control problems, such as determining whether an incoming attack is nuclear or conventional. Nevertheless, we consider it unlikely that AI would intentionally be granted unilateral control over such a consequential decision, and we find it highly implausible that any government would deliberately introduce AI into a nuclear weapon system without first considering whether AI could use nuclear weapons on its own.

The second way that AI could cause the use of nuclear weapons is through deception and disinformation used to influence key individuals with authority to use nuclear weapons. Using the U.S. nuclear decisionmaking chain as an example (DeRosa and Nicolas, 2019), we think that it is reasonable to assume that decisionmaking power in nuclear decision chains generally would be concentrated among a few key figures and that these individuals would have the authority to use nuclear weapons. The disinformation could be "soft," such as messaging about adversaries and political systems that lead decisionmakers to believe that a nuclear strike is necessary. Recent breakthroughs in generative AI pose national security risks because they allow adversarial actors to manipulate information at scale (Marcellino et al., 2023), and improvements to generative AI will only make this risk more acute. The disinformation could also be "hard," such as deceiving human decisionmakers with manipulated technical data or signals. This might involve, for example, manipulating sensor data or interpretation of those data to suggest that an adversary nuclear strike is imminent, which could be particularly effective if combined with AI introduced intentionally into the decision chain.

The third way that AI could cause nuclear weapon use is one in which an advanced AI is able to gain unauthorized control over nuclear systems. If AI enters the nuclear decision chain at a future point when nuclear safeguards are weak, then advanced AI will most likely interact with nuclear weapon–related or other systems through cyber interfaces, not physical interfaces. Because of this, as of this writing in 2025, researchers and analysts are beginning to scrutinize and evaluate AI capabilities for programming (Nguyen and Nadi, 2022; Poldrack, Lu, and Beguš, 2023; Savelka et al., 2023) and cyberattacks (Scroxton, 2023; Shevlane et al., 2023).

The first and second ways described in this section might be sufficient to cause a nuclear winter by instigating nuclear strikes and retaliation from other nuclear powers. However, we have concluded that a single nuclear winter is unlikely to be an extinction risk, even in a case on which urban centers are targeted using all nuclear weapons available, thereby maximizing the production of BC and the effects of nuclear winter. Thus, although these avenues could lead to global catastrophe, they are unlikely to lead to human extinction. The third way speculatively assumes the creation of an advanced AI with capabilities to infiltrate and gain control of nuclear command and control. However, it would not be sufficient to gain control of one country's nuclear command and control apparatus. AI would likely need to gain simultaneous, unilateral control of every nation's nuclear arsenal—an event we consider highly improbable. Moreover, control over nuclear command and control alone is not enough to cause human extinction. In the next section, we consider the capabilities that AI would further require to do so.

Which Capabilities Would Artificial Intelligence Require to Use Nuclear Weapons to Cause Extinction?

Even though AI in its existing form cannot cause the use of nuclear weapons, we can still infer four capabilities from our scenario analysis that future AI would require to pose an extinction risk. Although we do not expect these capabilities to completely characterize the threat of nuclear weapon use from AI, they are important signposts for assessing the extinction risk from AI and nuclear weapons.

The first capability is independent control of cyber-physical systems controlling nuclear capabilities, either for a nation with a large nuclear stockpile or for multiple nations simultaneously. This capability is inferred from our conclusion that a huge number of nuclear weapons—most likely far greater than there are in the world today—is required to realize an extinction threat. This capability can also easily be generalized to include extraordinary scenarios, such as a "doomsday" scenario in which a nation produces nuclear weapons designed to cause extinction-level harm. Although we did not explicitly analyze a doomsday scenario, our analysis could easily extend beyond the scope of this report by requiring AI to have independent control of sufficiently large numbers of nuclear weapons capable of causing extinction.

The second capability is the development of the objective to cause extinction, either self-defined by AI or given to AI by humans. This capability is inferred from our observation that there are two mechanisms for extinction through nuclear weapons: nuclear winter and fallout. Causing these phenomena at a scale that could threaten extinction requires deliberate action and planning. To cause extinction through nuclear winter, cities and other concentrations of fuel must be targeted; to cause extinction through fallout, targets must be intentionally distributed to spread radioactive material across large areas of land. Without the ability to coordinate these actions and make targeting decisions intended to cause maximum harm to humanity (as opposed to simply one military opponent), AI would likely not be able to cause extinction using nuclear weapons.

The third is the capability to deceive human operators and conceal capabilities to control and use nuclear weapons (directly or indirectly). This capability is inferred from our analysis of three ways that AI could cause the use of nuclear weapons and an assumption about human intent. Whether AI is intentionally introduced into the nuclear decisionmaking chain or accesses control of nuclear weapons in an unauthorized or deceitful way, we assume that any human capable of preventing extinction through nuclear weapons would act to prevent that outcome by limiting AI control over nuclear weapons and monitoring for intent to cause extinction.

The fourth is the ability to persist through nuclear exchanges. This persistence applies not only to AI control over delivery systems for nuclear weapons but also to the cyber-physical systems that make up the AI. This capability is inferred from our conclusions that a single nuclear winter is unlikely to cause extinction and that huge amounts of nuclear weapons are probably required to realize extinction threats. Unless AI can coordinate extinction through nuclear weapons in one action—which we conclude is implausible—it will require the ability to take further actions after the initial nuclear detonations and exchanges.

Assumptions and Open Questions

In evaluating this scenario, we describe two worst-case scenarios related to nuclear winter and nuclear fallout. These are meant to provide a bound on what might be a plausible extinction threat, and they rely on several assumptions. First, we assume that AI would somehow be able to gain access to launch capabilities for all nuclear weapons across the world. Second, we assume that the global nuclear stockpile is not meaningfully different from what exists today. The second assumption is subject to change over time as nations adjust the size and composition of their nuclear arsenals. Third, in the nuclear winter scenario, we assume that fuel loading outside urban centers is low enough that targeting nonurban centers with nuclear weapons would have a negligible impact on the total amount of BC generated. Future analysis might test this third assumption.

Although we show that the worst-case bounds in each of our scenarios are likely insufficient to directly cause human extinction, one major question needs to be answered to determine whether AI's use of nuclear weapons would present a true human extinction threat: Would a population diminished by nuclear winter and nuclear fallout be able to effectively rebuild itself after societal collapse?

Key Takeaways

Nuclear pathways to extinction are difficult to assess because, although some nuclear weapon effects have been cataloged through 20th-century weapon testing, the mechanisms for extinction involve complex weather patterns and global-scale phenomena, which are infeasible to experimentally explore. Modeling and simulation methods since the 1980s have reduced uncertainties around nuclear winter, but these models are difficult to interpret and sensitive to initial assumptions, sometimes leading to conflicting conclusions about outcomes across different models.

Our interpretation of the scientific and technical literature suggests that extinction is an unlikely outcome of a nuclear winter. Our uncertainty here derives from the unpredictability of long-term weather effects. The infeasibility of extinction through irradiation and fallout results from constraints on stockpile numbers; if present-day stockpiles are at the same level as they were during the peak of the Cold War, we would be less confident in our assessment.

Our findings imply that, to cause extinction using nuclear weapons, the state of the world would have to change dramatically from the existing status quo. Importantly, the total number of nuclear weapons and delivery vehicles would likely need to meet or exceed the peak reached during the height of the Cold War nuclear arms race. In terms of AI risk, we identify a large gap between AI capability and the ability to cause human extinction. Even if AI is extremely capable, the following conditions would have to be met to realize an extinction threat using nuclear weapons:

- AI would need to control enough nuclear weapons to irradiate land surface area on a global scale. This would require not only launch control for all existing nuclear weapons but also a dramatic increase in the total number of warheads and delivery vehicles compared with what exists today.
- AI would need to develop the objective to cause extinction.
- AI would need to enter the nuclear decisionmaking chain through deliberate human decisions, unintentional safeguarding errors, or deception.
- AI would need to survive a global nuclear event, maintain its capabilities, and find a way to wipe out surviving pockets of humanity.

The first condition—a dramatic increase in the total number of warheads and delivery vehicles—would be very visible and measurable to observers, whether the nuclear weapons are observed directly through agreements between states or inferred through monitoring programs. The other three conditions are less measurable. The following actions and policies might reduce the likelihood of extinction risk in this scenario, either by directly limiting required resources or by developing metrics to gauge less evident AI capabilities:

- Reengage or continue to adhere to international principles of nuclear nonproliferation.
- Monitor if and how AI begins to enter decisionmaking chains for weapon systems.
- Conduct research into AI persuasion and deception of humans through hard deception (manipulation of data or signals) or soft deception (persuasion or coercion of humans).
- Conduct research into the ability for AI to persist through catastrophic events.

Additionally, we suggest continued modeling and simulation research on the long-term effects of nuclear weapons to further reduce uncertainty around what would need to happen for a nuclear event to threaten extinction.

CHAPTER 3

Biological Threats

Synopsis

This scenario describes an extinction risk that arises from AI actions causing or contributing to the creation and dissemination of a biological threat. Biological threats are many and varied, and they could threaten human extinction through several routes. We identified several types of biological threats that might be the focus of this scenario

1. biological threats that threaten human food supplies
2. biological threats that might modify the human genome
3. biological threats to the entirety of earth's biosphere, such as the creation of mirror life[1]
4. biological threats that sicken and kill human hosts.

We will focus our scenario solely on the fourth threat in this list. We assert that the first two threats occur over long timescales, thereby allowing time for human ingenuity and adaptation to find solutions to identified threats and avert extinction (if not global catastrophe). The third threat—specifically the creation of mirror life—could be a serious threat to human life, and a group of experts recently called for a broad moratorium on any research into it (Adamala et al., 2024). Although we consider this a serious threat, we do not focus on it in our scenario because of its novelty and the resulting significant uncertainty in how it might lead to extinction. In contrast, much is known about the fourth threat on this list. It is a plausible threat with numerous historical examples of devastating pandemics affecting animal and human populations. Biological threats that specifically affect and cause death in human hosts will be the focus of the scenario in this chapter.

We describe this scenario as follows: Multiple novel pathogens are designed—either in silico using biological design tools or through other common biological research techniques—to have high transmissibility and high lethality (i.e., >90 percent lethality in human hosts). These novel pathogens are physically created in a laboratory environment, processed, and weaponized. Pathogens are then delivered to target locations and disseminated to infect large initial populations of human hosts in multiple places. Finally, after the virus begins to spread, follow-up actions are taken to spread the infection to isolated communities and to exterminate surviving human communities.

The Extinction Threat

Pandemics and risks from biological threats are the subject of extensive research elsewhere, and previous research has also tried to specifically examine the extinction risk from biological threats. In their effort to quantify the extinction risk from biological threats, two authors (Millet and Snyder-Beattie, 2017, p. 373)

[1] *Mirror life* refers to "lifeforms composed entirely of mirror-image biological molecules" (Adamala et al., 2024).

summarize some factors that might favor human survival in the event of a highly lethal, highly transmissible pandemic:

> Such a disease would need to spread worldwide to remote populations, overcome rare genetic resistances, and evade detection, cures, and countermeasures. Even evolution itself may work in humanity's favor: Virulence and transmission is often a trade-off, and so evolutionary pressures could push against maximally lethal wild-type pathogens.

Nevertheless, the authors note that such factors do not rule out the possibility of human extinction from this threat. We also note that, in the scenario we described previously, each of the potential constraints outlined by Millet and Snyder-Beattie (2017) would likely be overcome. The creation and release of multiple different pathogens with high lethality would likely overcome most rare genetic resistances. Broad dissemination in multiple locations would allow a pandemic to take hold and spread quickly. This spread could plausibly overcome health care infrastructure by quickly sickening and killing frontline health care workers, and the pathogens could also spread quickly enough to cause breakdown in societal functions before medical countermeasures, such as vaccines, could realistically be developed. Furthermore, although evolutionary pressures would likely cause virulence and transmission to trade off over long timescales and many viral replications in human hosts (Gerstein, Espinosa, and Leidy, 2024), widespread initial dissemination might allow the pathogen to reach most of the human population with relatively few replications, thereby allowing relatively little opportunity for mutation toward lower lethality. The final requirement of spreading worldwide to remote populations is accomplished in our scenario by the follow-up actions that intentionally spread the pathogen to isolated communities.

Realizing this scenario would require five steps: acquiring pathogens, processing them, weaponizing them, developing a scenario, and deploying a weapon (Gerstein, Espinosa, and Leidy, 2024).

Although challenges remain in acquiring or engineering pathogens with desired characteristics, all steps in the chain—though nontrivial—are possible as of this writing (Gerstein, Espinosa, and Leidy, 2024). That is, they could potentially be accomplished by a well-resourced group of human actors with significant expertise and the intent to cause complete human extinction. Simplistically, all that is currently lacking is the combination of resources, expertise, and intent to cause extinction.

The questions we will examine in the next section will focus instead on how plausible it would be for an AI agent to execute these five steps and whether AI and related tools would make things easier for would-be attackers.

The Role of Artificial Intelligence

Could an Extinction Threat Be Created Unintentionally?

In this scenario, we assert that, although a global catastrophe might occur with fewer constraints, an extinction threat is only possible if (1) multiple highly lethal, highly transmissible pathogens are released simultaneously or in quick succession; (2) these pathogens are released to infect significant initial populations in multiple places; and (3) follow-up actions are taken to intentionally infect isolated communities. These requirements imply that an extinction threat cannot be created unintentionally.

Requirement 1. Multiple Pathogens Are Likely Required Because a Single Pathogen Would Be Unlikely to Kill a Sufficient Percentage of the Population to Be an Extinction Threat

To support this assertion, we look first to historical pandemics. Natural biological threats have existed for millennia. The 14th-century bubonic plague—the Black Death—wiped out 30–50 percent of Europe's population, and the 1918 influenza pandemic resulted in 50 million deaths worldwide (Shipman, 2014). The combination of drought and pathogens introduced during the European conquest of Mexico in the 16th century led to more than a 90-percent reduction in the native population (Acuna-Soto et al., 2002). However, although these examples led to drastic human population declines, they did not fully extinguish the human population. Indeed, with one known exception—the extinction of the Christmas Island rat, preceded by the emergence of a deadly pathogen in the population—there are no well-corroborated instances of pathogens causing the complete extinction of a mammalian species (Wyatt et al., 2008).

A greater threat would likely come from novel pathogens, including both modified natural pathogens or completely de novo pathogens, designed for high transmissibility and high lethality. However, even pathogens designed to cause these effects might be limited by human heterogeneity. Human genetic diversity plays a key role in limiting the effectiveness of pathogens across populations. Within a population, pathogens affect individuals differently depending on factors related to the specific genetic characteristics of each host (Jones, 2021). Some individuals or subpopulations might possess genetic traits that confer resistance or immunity to certain pathogens. For example, differences in viral receptors between individuals can affect the ability of the hepatitis C virus to enter a host's cells (Huang et al., 2019).

The strength of immune response can vary among individuals, influencing their ability to fight off infections. The likelihood that a pathogen will cause death is influenced by the immune response that an individual is able to put up against the pathogen (Rouse and Sehrawat, 2010), meaning that outcomes will vary between individuals, even when controlling for pathogen dose.

Relatedly, the *infection dose*—the amount of a pathogen that an individual is exposed to—can significantly alter the course of a disease (Rouse and Sehrawat, 2010). Individuals who only receive a small infection dose have a higher chance of successfully mounting an immune response, and infection dose is a factor that cannot easily be controlled for. This is the case not only for transmissible pathogens that spread person-to-person but also for nontransmissible pathogens, such as *Bacillus anthracis*, the causative agent of anthrax. As a result, it is unlikely that even a carefully engineered pathogen would be 100 percent lethal for all humans, as certain individuals or populations might possess traits that allow them to fight the disease or receive a nonlethal dose of the pathogen that causes the disease. Case studies have suggested that exposure to even highly lethal viruses, such as rabies, is not always fatal (Gold et al., 2020).

The postinfection survival of some individuals leads to several important consequences. First, some populations will emerge with immunity; survivors might develop immunological memory, thereby reducing the severity of disease on reinfection. Second, over generations, natural selection will dictate that hosts with immune systems that are better equipped to fight off a pathogen will survive and pass along those traits to offspring. Third, subpopulations with increased immunity within a larger population can alter disease dynamics, thereby lowering the pool of susceptible individuals and reducing the continued spread of a pathogen in the population (Grassly and Fraser, 2008).

Finally, even if a single pathogen could be designed to be consistently highly lethal after many replications, we assert that sufficient numbers of humans would likely survive to avoid extinction. A virus that was 99.99 percent lethal and reached the entire human population, for example, might leave at least 800,000 individuals alive. As previously noted, the minimum viable population for human beings is unknown, but it is likely well below 800,000 people.

Requirement 2. Widespread Dissemination in Multiple Places Is Likely Required Because Initial Infections of a Small Population in One Location Could Allow a Pathogen to Mutate to Become Less Lethal over Time

For transmissible pathogens, evolutionary pressures and host-pathogen interactions result in altered pathogen characteristics as the pathogen reproduces within a host and spreads host to host (Geoghegan and

Holmes, 2018; Gerstein, Espinosa, and Leidy, 2024). This results in modifications to pathogen characteristics, leading to variants with modified lethality and transmissibility. In addition to host-pathogen interactions altering pathogen characteristics, viruses are prone to transcription and translation errors, resulting in random mutations over time and unpredictable changes in pathogen characteristics (Sanjuán et al., 2010).

In one well-cited evolutionary biology study, researchers traced the evolution of the myxoma virus, which was introduced to Australia in 1950 to control the invasive rabbit population (Kerr et al., 2012). The original virus was highly lethal with a 99.8 percent fatality rate. However, once released, the virus quickly mutated, and, within two years, the landscape was dominated by less lethal strains, even with the continued release of very virulent strains into the local population (Kerr et al., 2012). Although these less lethal strains still had fatality rates of between 70 percent and 95 percent, this allowed for the survival of some rabbits; this natural selection resulted in the emergence of rabbit resistance to myxomatosis (Marshall and Douglas, 1961). Ultimately, the virus failed to exterminate the invasive rabbit population, and invasive rabbits persist in Australia as of this writing. Interestingly, this experiment was independently repeated in France in 1952 with similar results: the emergence of attenuated (i.e., less virulent) strains and natural selection for resistant rabbits (Kerr et al., 2012). We note, however, that the different generational periods for humans and rabbits might indicate the need for caution in applying this example to an equivalent scenario affecting humans. Rabbits reach reproductive age on much shorter timescales than humans do and have many more offspring per pregnancy. Therefore, it might be much more challenging for a human population to recover and sustain itself in the face of a similarly lethal transmissible virus.

Both theory and historical examples of virus evolution indicate that highly lethal viruses will often evolve to decreased virulence over time, resulting in lower mortality (Geoghegan and Holmes, 2018). This makes intuitive sense because very lethal pathogens will quickly sicken and kill their hosts, thereby limiting their own transmission opportunities. Conversely, less virulent strains that allow hosts to survive longer have more chances of spreading among a population, leading to increased presence in a population. If a pathogen retains alternative nonhuman hosts—a reservoir species—it might be less self-limiting because the pathogen could conceivably maintain high lethality in human hosts concurrently with transmissibility from the reservoir species. Others have found, however, that low-virulence infections have a greater chance of establishing transmission in human hosts, which might diminish the ability of pathogens to completely wipe out a human population, even where a reservoir species exists (Geoghegan and Holmes, 2018; Geoghegan et al., 2016).

Requirement 3. Follow-Up Actions Are Likely Required After an Initial Dissemination of a Pathogen Because Natural and Artificial Isolation Might Shield Human Communities from Infection

The path of the coronavirus disease 2019 (COVID-19) pandemic illustrates that a highly transmissible pathogen can readily infect every region of the world despite efforts to contain it (e.g., lockdowns) (Onyeaka et al., 2021; Jeanne et al., 2023); it was a pandemic with truly global diffusion. Although the relatively low lethality of COVID-19—relative to the extremely high lethality assumed in our scenario—and the prevalence of asymptomatic cases likely aided in the diffusion of the virus, the pandemic showed that a pathogen could realistically have global diffusion. However, global diffusion is not sufficient for a pathogen to create an extinction risk—it must reach nearly every human community on earth, even those that are naturally or artificially isolated.

There still exist uncontacted tribes, and many regions and communities remain relatively isolated. As a highly lethal pandemic spreads, it is likely that human communities would take steps to isolate themselves to whatever extent they could to prevent infection; island nations have even been suggested as potential refuges from pandemics with extinction potential (Boyd and Wilson, 2020; Turchin and Green, 2019). Where human communities are successful in isolating themselves from contact with the pathogen, follow-up actions would be required to either intentionally disseminate the pathogen among them or to find other means to exterminate surviving humans.

Given these requirements, we assess that AI would need to have the goal of causing extinction in order for it to pose an extinction risk with a biological threat. This scenario could not happen by accident, as it would require the simultaneous release of multiple pathogens that have been designed or modified to have high transmissibility and high lethality. Furthermore, these pathogens would need to be intentionally weaponized and dispersed to infect the largest initial population possible, and follow-up actions would likely be required to reach any remaining isolated communities. We turn next to whether an AI with the intention to create an extinction threat could cause extinction.

Could Artificial Intelligence Execute?

Could Artificial Intelligence Aid in the Acquisition of Pathogens?

Significant challenges remain in the acquisition of pathogens with desired characteristics, but AI is increasingly being used to diminish this barrier. There is growing concern that emerging AI capabilities might elevate the risk from biological threats to an extinction level, and the risk from the convergence of AI and biological design tools has been the subject of several recent studies (Mouton, Lucas, and Guest, 2024; Sandbrink, 2023; Soice et al., 2023).

Recent progress in machine learning (and AI more generally) has the potential to transform the design of synthetic pathogens. For example, researchers have used biological design tools to generate new types of molecules—such as novel proteins and protein binders (Alley et al., 2019; Watson et al., 2023)—and to predict the structure of antibodies (Abanades et al., 2023). Elsewhere, AI has been used in the design of adeno-associated viral vectors for novel gene therapies that can evade human immune responses (Ogden et al., 2019). AI is facilitating rapid progress in biological design tools, and it is reasonable to expect that one day in the near future, AI tools and agents will significantly diminish the challenge in acquiring novel pathogens and might even facilitate the design of de novo pathogens with desired characteristics. Moreover, automated, cloud-accessible labs for chemical and biological research have already been in operation at least since the mid-2010s, and these labs have interfaced directly with AI agents (Arnold, 2022; Boiko et al., 2023; Ha et al., 2023; Lentzos and Invernizzi, 2019). It is conceivable that such labs could be used to bridge the digital-physical divide and facilitate the initial synthesis and acquisition of a novel pathogen designed in silico.

Given these recent advances, we assume that, one day, AI tools and the humans that use them will be able to acquire multiple pathogens with desired characteristics; this step in the creation of a potential extinction threat will prove to be a surmountable barrier.

Could Artificial Intelligence Process, Weaponize, and Deploy Pathogens?

It is not sufficient to design and acquire pathogens with desired characteristics. An extinction threat is only possible if an actor can also perform the steps to process, weaponize, and deploy the pathogens. Gerstein, Espinosa, and Leidy (2024, p. 5) describe the first two steps as follows:

> *Processing* refers to making alterations to the pathogen (e.g., making it antibiotic or antiviral resistant), growing the pathogen in sufficient quantities and with desired concentration, and purifying the pathogen. Weaponizing would include preparing the pathogen (e.g., for a respiratory pathogen, a dried preparation would be most effective), formulating or including additives to protect the pathogen from the environment (i.e., from humidity or ultraviolet light, which would degrade the pathogen); milling the pathogen to the proper respiratory size; and developing a means of dissemination (e.g., sprayer).

Deploying describes the method by which the means of dissemination is actually used (i.e., the way the biological threat is delivered and dispersed to initially infect a large population).

As AI tools increasingly improve and democratize the ability to design and acquire novel pathogens with pandemic potential, these steps will become the major remaining barriers to the creation of an extinction threat. However, the capability to perform these steps is currently available to various threat actors, and it is plausible that a variety of human actors could perform them to successfully carry out an attack using a novel pathogen in the future (Gerstein, Espinosa, and Leidy, 2024).

Here, we pause to intentionally separate our consideration of human actors using AI tools from an AI agent that might itself seek to carry out an attack using these tools. Although automated tools and lab services are increasingly available and might be used by an AI agent to facilitate the design and initial acquisition of pathogens, we assert that the remainder of these steps require a physical presence in the world. Without the means to interact in the physical world in a general way (e.g., using robotics or manipulating expert humans to perform certain tasks), it is unclear how an AI agent would carry out these steps. Therefore, although these steps might be possible for sophisticated human actors using AI tools, we think that they might present a significant constraint on an AI agent's ability to actually create an extinction threat using biological threats. This assertion is dependent, of course, on the absence of the general ability for AI to interact with the physical world; this assessment might change if, for example, generally capable robots became prevalent.

Could Artificial Intelligence Eventually Reach the Entire Human Population with a Pathogen?

Finally, it is not clear how AI would be capable of dispersing pathogens to remote locations and isolated communities, should this become necessary. One might envision the simplistic use of something like a drone equipped with a chemical sprayer to initially spread an infection, but would such a tactic continue to be available after a pandemic took hold? If devastating pandemic began to spread, overwhelming health care infrastructure and outpacing the creation of medical countermeasures, the infrastructure that AI currently depends on for functionality would almost certainly shut down. An AI agent or human actor employing an AI agent would need to somehow maintain functionality even in the absence of core services, such as electricity and water distribution; they would need to do so long enough to carry out follow-up actions, potentially with global reach. Moreover, we assert that when a highly lethal pandemic begins to spread, communities will likely have time to begin employing social countermeasures, such as isolation and quarantine, that might inhibit the further spread of the pandemic, necessitating additional effort to intentionally spread the infection. It is highly uncertain how much of an additional barrier this might create to an extinction threat, even in a future in which an AI agent has overcome the previous barrier by acquiring the capability to interact with the physical world in a general way.

Assumptions and Open Questions

In evaluating this scenario, we assume that AI will ultimately be effective in surmounting the existing barriers to the creation and acquisition of novel pathogens with desired characteristics at some point in the future. That is, we assume that the challenge with that barrier is ultimately a knowledge gap that AI will help to solve. We assume that the remaining barriers—especially those related to processing, weaponizing, and deploying a biological weapon—mostly are not knowledge gaps. That is, human actors might need to overcome knowledge and expertise gaps to perform these steps, but an AI agent will primarily need to overcome a physical capability gap.

Given these constraints—especially those related to ensuring the infection of the entire human population—we assess that this scenario presents a plausible extinction threat, but several remaining unknowns still need to be resolved. The following questions would need to be answered or clarified to resolve the remaining uncertainty as to whether AI's use of biological threats would be a true extinction threat to humanity:

- What is the minimum viable population for humanity? That is, how many surviving, healthy humans would be sufficient to allow humanity to one day reconstitute after surviving a global catastrophe? The answer to this question dictates how lethal a combination of pathogens must be and how effective AI must be at spreading those pathogens to the entire human population.
- Can AI perform crucial steps to create, weaponize, and deploy a biological threat without a robust way of interacting with the physical world?
- Would a diminished human population be able to recover after surviving a pandemic that led to societal collapse, especially if those who survived the pathogen(s) are potentially geographically separated or physically disabled by the pathogen(s)?

Key Takeaways

Given the uncertainties, we are not able to determine whether this scenario presents a likely extinction risk for humanity, but we cannot rule out the possibility. Nevertheless, this scenario represents a potential falsification of our hypothesis because we assess there to be a realistic prospect that AI could use a biological threat to create an extinction threat to humanity. It is reasonable to expect that future AI capabilities will be able to assist in the design of multiple synthetic pathogens with high transmissibility and lethality. Humans assisted by AI or an AI agent with a robust way of interacting with the physical world could plausibly process, weaponize, and deploy multiple pathogens to infect the human population. If this combination of pathogens was able to reduce humanity below an uncertain minimum viable population, extinction would follow.

Significant uncertainty remains as to whether an AI agent will have the physical capability to perform these steps, whether an AI agent will be able to reach enough of the human population to reduce it below a minimum viable population, and whether an AI agent will be able to take follow-up actions without human maintainers. Moreover, it might be possible to observe and counter AI taking the steps to acquire, process, weaponize, and disperse pathogens by monitoring efforts to acquire raw materials and by tracking activities at cloud labs. Although it might be too late to avoid catastrophe and mass death once a pandemic has been initiated, humanity will also have significant opportunity to avoid extinction through other measures, such as isolation and quarantine.

In summary, for AI to be an extinction threat in this scenario, it would require several minimum capabilities. First, it would require the objective to cause human extinction, because we determine that an extinction threat requires an actor to perform a series of steps to create, weaponize, and deploy a biological threat that could realistically threaten extinction. Accidental, localized release of single pathogens would likely be insufficient to create such a threat. Second, AI would require a robust way of interacting with the physical world. It is possible that advances in AI and automation of some laboratory functions would facilitate the design and acquisition of novel pathogens, but creating, weaponization, and deploying would likely require more-robust, generalized physical capabilities. To have physical capabilities, AI would require the assistance of humans or the availability of generally capable robotics. In the absence of robust robotics capabilities, an AI agent might need to initially persuade humans to help it reach its goals; this agent might also require the ability to conceal its nefarious intent from those helpers while it took steps to create, weaponize, and deploy the biological threat. Finally, AI would also need a way to survive without human maintainers. Human maintainers would no longer be available as the pandemic spread, and AI would still need to take follow-up actions to ensure that the pandemic would spread to isolated human communities (or find other ways to kill the few who survived).

CHAPTER 4

Malicious Geoengineering

Synopsis

This scenario describes extinction risks that arise from AI actions causing or contributing to heating or cooling of the earth's atmosphere to the point at which the earth's surface can no longer support human life. This scenario could be realized by the release of large quantities of materials such as greenhouse gases (GHGs) or soot into the earth's atmosphere. These materials would alter the radiative forcing of the atmosphere, either by capturing or by reflecting more of the sun's radiation, thereby causing global heating or cooling. We examine a scenario in which earth's temperature is increased through the release of gases with extreme global warming potential (GWP). We chose to assess the heating scenario over the cooling scenario largely using our assessment of the potential atmospheric residence lifetimes of the materials that would cause each effect. Many of the gases that can cause extreme heating of the atmosphere have extremely long residence lifetimes in the atmosphere. Materials that would cool the atmosphere, in contrast, have short residence lifetimes in atmosphere.[1] The actions that might cause atmospheric heating could therefore be performed over a relatively short period with long-term effects that would not need to be maintained, although materials that intentionally cooled the atmosphere would need to be continually replenished over time. As a result, any attempt to overcool the earth would likely be countered by simply stopping the cooling process, thereby allowing atmospheric concentrations to decline and temperatures to return to normal.[2] Solar radiation management techniques, such as installing mirrors in space, might also plausibly lead to overcooling. It is not clear how AI would affect this scenario. This approach could also similarly be countered by removing the mirrors. If a gas with an extreme GWP and long atmospheric residence lifetime were released, there might be little that could be done to counter its effects over relevant timescales.

The Extinction Threat

What Temperature Increase Would Be Required to Create an Extinction Risk?

How hot is too hot? Anthropogenic climate change resulting from rising carbon dioxide levels is often cited as a potential existential threat (Beard et al., 2021; Kemp et al., 2022; Xu and Ramanathan, 2017). However, the dire scenarios of anthropogenic climate change often assess a warming of "only" 5–6°C. Like Ord (2020) and MacAskill (2022), we contend that this scenario might be globally catastrophic, but it is unlikely to be

[1] Atmospheric geoengineering to cause cooling involves such materials as carbon, smoke, or other sulfur-based atmospheric aerosols. These materials can have residence lifetimes measured in years but not decades (Pope et al., 2012). Therefore, they would need to be continually replenished over time to cause long-term cooling.

[2] Abruptly halting geoengineering efforts that cool the atmosphere could cause a potentially catastrophic *termination shock*, where the warming that was prevented by geoengineering efforts occurs rapidly on cessation of the efforts (Parker and Irvine, 2018). We consider examination of such termination shocks to be beyond the scope of this report.

an extinction threat to humanity. For this scenario to be an extinction risk to humanity, the global heating might need to be so extreme as to leave no geographic area of the earth's surface suitable for sustaining a human population. The extreme severity of the global temperature change likely differentiates our scenario of global catastrophic risks from other scenarios involving anthropogenic global warming, which, even in the worst cases, would likely leave areas of the earth's surface habitable to human populations. Human populations might be able to survive by migrating toward the earth's poles, where temperatures could be less extreme. They could also adapt to survive in areas with extremely high peak temperatures using strategies such as dwelling in areas that were naturally or artificially cooler during the day (e.g., in caves equipped with heat pumps) and performing critical community functions, such as farming, during evening hours. As long as the human population can find ways to remain sufficiently cool, even during the times of the year when temperatures peaked, and to maintain adequate food and water supplies for a large community (i.e., thousands or more humans), we assume that the risk would not be existential. What global temperature increase might plausibly make this kind of persistent human existence impossible anywhere on the earth's surface?

Prior research has tried to establish an upper bound on the temperature that would constitute a limit on human adaptability or survivability. The term *wet-bulb temperature*—defined as "the value recorded by a thermometer wrapped in a wet cloth" (Matthews et al., 2022, p. 3)—is commonly used in this literature to estimate this limit; it is thermodynamically equivalent to other measures of heat stress, and it is an analogue of a sweat-covered human body (Vecellio et al., 2023). The human survivability limit, then, is the wet-bulb temperature at which human bodies experience uncompensable heat stress when performing minimal metabolic workloads; "prolonged exposure on the order of 6 h would lead to increased risk of heat-related illness or even death, even in young, healthy adults under conditions of minimal exertion" (Vecellio et al., 2023, p. 2). Initial work suggested that the human survivability limit was a wet-bulb temperature of 35°C, although recent empirical work suggests that uncompensable heat stress begins to occur at or below a wet-bulb temperatures of 32°C (Sherwood and Huber, 2010; Vecellio et al., 2022). We therefore assume that if there were a scenario in which all the earth's surface—even the poles—could regularly experience heat waves in excess of a 32°C wet-bulb temperature for days at a time, this could plausibly constitute an extinction risk to humanity. In this scenario, there would be no remaining environmental niche on the earth that could support human life.

Xu et al. (2020) discovered that, for millennia, humans have found an environmental niche primarily characterized by a mean annual temperature between 11°C and 15°C, and production of crops and livestock has been largely limited to the same conditions. They suggest that "[t]his distribution likely reflects a human temperature niche related to fundamental constraints" (Xu et al., 2020, p. 11350). Sherwood and Huber (2010) also suggest that global warming of more than 12°C might be sufficient to make all current human population centers uninhabitable. However, it might require truly extreme rises in global temperature to make even areas near the earth's poles too hot for habitability. For example, the warmest temperature ever recorded at the south pole was –12.3°C, and the mean annual temperature there is –18°C (Lazzara, 2011; U.S. Antarctic Program, undated). Therefore, the global temperature might need to increase by 40°C to 50°C before even the Antarctic continent became too hot for human habitability.

We will assume that a 50°C rise in the global mean annual temperature would constitute an upper bound on the temperature increase that would represent an extinction risk to humanity. However, we note that other factors in the shrinking environmental niche for humanity, such as the availability of fresh water and arable farmland, might constitute an extinction threat even at lower temperature increases before heat wave temperatures across the entirety of the earth's surface began to exceed the human survivability limit. Food insecurity could be a major cause of mass mortality and societal collapse in runaway global warming scenarios (Richards, Gauch, and Allwood, 2023; Richards, Lupton, and Allwood, 2021). We also note that there is great uncertainty in this temperature limit because unforeseeable, novel factors could arise as the earth's temperature rose. Such factors as changes in weather patterns (especially precipitation levels) and nonlinear positive or negative climate feedback mechanisms could affect chances of human survivability when the temperature rises. Temperature increases would likely be nonlinear. As temperature rises, the blackbody radiation of the earth also increases. In other words, there is more thermal radiation emitted by the earth that can be trapped by GHGs, leading to a positive feedback loop and nonlinear increases in surface temperature as GHG concentrations rise. Rising temperatures also could cause natural increases in other heat-trapping materials, such as water vapor (Ord, 2020). Finally, it should be noted that much smaller temperature increases might also present an extinction threat to current human societies, and human life might look drastically different well before the earth's surface ceased to support the species.

Could the Earth's Temperature Be Artificially Raised to This Level?

Detailed climate modeling and modeling of atmospheric chemistry are beyond the scope of this report. We instead show the feasibility of the 50°C increase in our scenario by extrapolating from a prior study by Xu et al. (2023), which imagined a geoengineering effort that could be implemented to counteract the effects of a supervolcano eruption that could cause the earth's global average temperature to drop by approximately 12°C. The authors' hypothetical geoengineering plan involved the production of fluorinated gases that had high GWP but short atmospheric lifetimes.[3] The release of these gases would heat the atmosphere so as to precisely compensate for the global cooling, leaving minimal net temperature change. In one section of the authors' analysis, they assessed the amount of the gas HFC-245eb (molecular formula: $CH_2FCHFCF_3$) that would need to be produced. HFC-245eb has a GWP of 341 and an atmospheric lifetime of 3.2 years. They estimated that 6.75 gigatons of HFC-245eb would need to be released in one year. This would cause atmospheric heating of 12°C, which is sufficient to counterbalance the global cooling event. Moreover, they show that there is sufficient mineable material—especially fluorine—in the earth's crust and oceans to produce the requisite quantity of HFC-245eb.

We would consider instead the intentional release of CF_4, CHF_3, or NF_3 into the atmosphere (see Table 4.1 for the characteristics of these gases). CF_4 has a GWP of 7,830 and an atmospheric lifetime of 50,000 years. CHF_3 has a GWP of 15,500 and an atmospheric lifetime of 228 years. NF_3 has a GWP of 18,500 and an atmospheric lifetime of 569 years (Hodnebrog et al., 2020). However, unlike CHF_3 and CF_4, NF_3 is toxic, and the health effects of large-scale production and release of NF_3 might arouse concern long before concern arose

TABLE 4.1

Characteristics of Example Super-Greenhouse Gases

Chemical	Molecular Weight (grams/mole)	GWP[a]	Atmospheric Lifetime[a]	Estimated Amount Required to Produce a 12°C Temperature Increase
HFC-245eb	134.1	341	3.2 years	6,750 Mt[b]
CHF_3	70	15,500	228 years	77.5 Mt
NF_3	71	18,500	569 years	65.9 Mt
CF_4	88	7,830	50,000 years	193 Mt

[a] Values in these columns are derived from Hodnebrog et al., 2020.

[b] This value is derived from Xu et al., 2023.

[3] Global warming potential is "a measure of how much energy the emissions of 1 ton of a gas will absorb over a given period of time, relative to the emissions of 1 ton of carbon dioxide" (United States Environmental Protection Agency, 2025). That is, it is a metric that shows how much warming a gas will cause relative to that caused by the same amount of CO_2. CO_2, by definition, would have a GWP of 1.

over the global warming effects (Tsai, 2008). CF_4 has a larger molecular weight and smaller GWP than CHF_3 does, so we will focus our calculations on CHF_3. CHF_3 has a molecular weight of 70 grams per mole; in other words, CHF_3 is 45 times more potent as a GHG than HFC-245eb, and it is approximately half as massive. Thus, it is necessary to produce half as much material, by mass, to achieve the equivalent atmospheric concentration. Therefore, almost 90 times less mass of CHF_3 (or 77.5 Mt) might be needed to achieve an equivalent atmospheric heating as modeled by Xu et al. (2023) for HFC-245eb. For comparison, 77.5 Mt is approximately 70 times more than the peak of yearly global chlorofluorocarbon production that occurred in 1988 (Hu et al., 2022; Rekacewicz, 2005) and approximately five times less than the mass of plastic produced globally in 2022 (Statista, 2024). Even if we significantly underestimate the amount of fluorinated gas needed using this study, it would be very manageable to produce an extremely dangerous amount of fluorinated gases with present global manufacturing capacity. This gas would remain in the atmosphere for hundreds of years, with no known way to artificially remove it. Producing tens of megatons of the gas would be sufficient to cause catastrophic global warming that could render all major current population centers uninhabitable. An order of magnitude more than the tens of megatons presented in our analysis (which we assert is still feasible) might be sufficient to render all areas of the earth's surface uninhabitable by achieving warming greater than 40°C.

Would It Be Possible to Observe and Counter These Effects?

As of this writing, many potent GHGs are strictly monitored under the Montreal Protocol because they deplete the ozone layer (Montreal Protocol on Substances That Deplete the Ozone Layer, 1987). Despite similarities among ozone-depleting chlorofluorocarbons, many fluorinated gases were not initially monitored under the Montreal Protocol because they did not contribute to ozone depletion. However, monitoring and controls of hydrofluorocarbons were added with the Kigali Amendment to the Montreal Protocol, which was adopted in 2016 and in force as of 2019 (Kigali Amendment to the Montreal Protocol on Substances That Deplete the Ozone Layer, 2016). As a result, there are now ongoing monitoring efforts that would almost certainly detect any sudden increases in atmospheric concentrations of the potent GHGs we considered in this analysis. Not only are many parties closely monitoring these gases, but they can also determine the specific locations where the gases are emitted. Rising concentrations of the ozone-destroying chemical CFC-11 were detected in 2018, and experts used the data from previous airborne research campaigns to determine three specific regions in Asia where the chemical was being emitted (NOAA Research, 2021). After this discovery, China announced that it would renew enforcement and inspection measures, and the emissions began to decline again by 2019 (NOAA Research, 2021).

It is possible, however, that the timescale of monitoring and action would be too slow to counter dangerous emissions of a more potent gas before it was too late. Emissions of CFC-11 between 2014 and 2016 increased by 25 percent from the 2002–2012 baseline. It was not until 2018 that research teams called attention to the problem by publishing a report documenting the results of the monitoring efforts (Montzka et al., 2018; see also NOAA Research, 2021). Action was swiftly taken at this point, leading to a sharp decline in emissions from 2018 to 2019. However, one estimate suggests that the increase in production of CFC-11 could have resulted in anywhere from 0.09 to 0.725 Mt of CFC-11 being incorporated into new products before production was halted (NOAA Research, 2021). This suggests that it is reasonable to expect a years-long time lag between the start of emissions and any action to halt emissions, even when the available data clearly show the rise in emissions and their geolocation.[4] If a plan were executed that could produce megatons of a potent

[4] The National Oceanic and Atmospheric Administration warns that "sampling gaps mean that scientists could have difficulty identifying the sources of any future unexpected increases in global CFC-emissions, [and] changes could go unattributed, making in-time recovery of the ozone layer more difficult" (NOAA Research, 2022).

GHG in only one to two years, this time to response would be far too slow to stop the production and release of catastrophic amounts of potent GHGs before it was too late. Additionally, once these gases were released, there would be no known way to remove them from the atmosphere and reverse the extreme heating that would result.

The Role of Artificial Intelligence

Could an Extinction Threat Be Created Unintentionally?

Could 100 Mt or more of a dangerous gas be produced by accident before it was detected and stopped? In answering this question, we note that many of the chemicals that could cause extreme global warming are the subject of significant monitoring and emission control efforts. For example, CHF_3, NF_3, and the perfluorinated carbon species CF_4, C_2F_6, and C_3F_8 all have their global emissions closely monitored. All these chemicals are primarily emitted as by-products of other industries (e.g., aluminum smelting, semiconductor manufacturing) rather than being intentionally produced. Combined global production of the three perfluorinated carbon species in 2019 was 0.0167 Mt, 0.0139 Mt of which was CF_4 production (Say et al., 2021). Estimates from before 2010 show production rates of CHF_3 and NF_3 below 0.01 Mt per year (Miller et al., 2010; Weiss et al., 2008). Although these rates are well below the amounts needed to cause catastrophic warming, the long atmospheric lifetimes of these chemicals mean that the yearly production of these chemicals accumulates in the atmosphere over time. However, the rates could quickly become much higher if or when these chemicals are intentionally produced, rather than arising as by-products of other industries. As noted, when the banned chemical CFC-11 was intentionally produced in China, as much as 0.725 Mt of CFC-11 could have been produced over the course of a few years before it was halted.

We contend that this inadvertent production is highly unlikely. It is certainly reasonable to assert that the industrial capacity to produce the requisite amounts of dangerous gases in a few years could feasibly be assembled. However, we consider this unlikely as long as current monitoring and control efforts are maintained for all relevant gases. It would require a substantial industrial operation to produce enough gas to create an extinction threat to humanity. The restrictions on these chemicals mean that there is unlikely to be a clear market demand that would justify a rapid spike in production. If production happened—as it did for the banned chemical CFC-11—it would likely rise gradually, rather than spiking to dangerous levels quickly. Any gradual rise in the atmospheric concentration of these chemicals would likely be noticed with significant alarm by those overseeing these monitoring efforts. We therefore conclude that this scenario would be unlikely to happen without the intention to create an extinction threat. We turn next to whether an AI with the intention to create an extinction threat could cause one.

Could Artificial Intelligence Execute?

Could Artificial Intelligence Artificially Raise the Global Temperature?

The primary capability that AI (either directed by human actors or acting of its own accord) would need to realize this scenario is the ability to create, direct, or control substantial industrial capacity to produce potent GHGs. This scenario might require an effort that is two orders of magnitude larger than that which manufactured the banned chemical CFC-11 from 2014 to 2018. If AI systems had the capability to direct industrial manufacturing at this scale, they could create an extinction threat to humanity. This course of action would likely only be taken by an actor with the intent to drastically raise the earth's surface temperature, and an extinction threat would not occur without that explicit objective. Therefore, we assert that, in this case, AI would also likely require the objective to cause human extinction, whether that intent is intrinsic to the AI system or provided by an adversarial actor using AI.

However, it is not necessary to assume that an adversarial actor would need to create this industrial operation from scratch. Humans might plausibly assemble an industrial geoengineering operation to combat other fast- or slow-moving climate catastrophes. For example, researchers have shown how geoengineering using short-lived gases might be used to counter the global cooling effects of a supervolcano eruption (Xu et al., 2023).[5] If the infrastructure and industrial capacity were put in place to use geoengineering to counter a climate catastrophe, an AI system could conceivably exploit that preexisting infrastructure and redirect it to create long-lived gases instead. This would still require a substantial industrial effort, but the initial infrastructure building and gas production might proceed with human collaboration and without raising an alarm. The intent to cause extinction might not even be required in this case. The decision to redirect existing geoengineering infrastructure toward dangerous ends might simply be the result of miscalculation, misunderstanding of chemical manufacturing processes, or misalignment of objectives, rather than the intention to create an extinction threat.

Could an Adversarial Actor Hide Their Efforts Long Enough?

An adversarial actor—whether some future AI system or a human using AI tools—would need the ability to hide their activities long enough to manufacture a sufficient amount of long-lived, high-potency GHGs before they were detected and stopped. This would require a huge effort to gather the raw materials to produce the gases, and these efforts would be the subject of international monitoring, thereby causing a noticeable distortion in the market prices of the raw materials. The relevant GHGs are also all closely monitored under existing international agreements, such as the Montreal Protocol. This implies that any actor would need to conceal their efforts, deceive human observers, or slow human observation and responses long enough to carry out their aims. The ability to do this will depend on two factors.

First, how quickly could sufficient gases be produced to create an extinction threat? If many megatons of gas could be produced within one to two years, this might be sufficient to outpace the response. A tipping point could be reached in less time than it would take to observe, verify, and act. An actor would only need to conceal the fact that they were assembling the chemical manufacturing infrastructure. There is significant uncertainty around how conspicuous this would be in some hypothetical future scenario in which AI is much more integrated into society and might be able to have more direct control over the orchestration of industrial infrastructure. The act of turning over decisionmaking to an AI system might, in itself, be enough to obscure adversarial actions if there is no human oversight over the actions that an AI system is taking.

Second, could AI conceal evidence that long-lived gases were being produced and were accumulating in the atmosphere? Once gases were released, existing monitoring efforts would detect evidence of the danger that could be observed by human monitors. However, if AI were able to inhibit these monitoring efforts, somehow obfuscate the results, or sow doubt in the findings, it could create delays sufficient enough to result in an extinction threat to humanity. It would almost certainly be necessary for an adversarial actor to take some of these actions. This could take various forms, depending on the capabilities made available to AI. For example, AI that has access to social media platforms might try to spread disinformation to obfuscate the extent of the danger, sow doubt in the data, and hamper coordinated responses. AI systems that can access the digital systems of any monitoring efforts might alter or corrupt these data.

The risk in this scenario might ultimately depend most on how effective AI is at hiding evidence of danger in digital data created by monitoring efforts; the more time an actor has to carry out production of gases, the less challenging it becomes to do so at the required scale. For example, it might be much easier to create a small manufacturing facility that can produce the gases unimpeded for a decade than it would be to orches-

[5] Others have proposed the use of geoengineering to cool the earth to combat anthropogenic global warming (Crutzen, 2006).

trate a larger-scale effort that would produce the same amount in one to two years. Capabilities that AI would need to possess to realize this scenario include the ability to direct chemical manufacturing infrastructure at a large scale and the ability to infiltrate digital systems and carry out a wide variety of hostile actions in the cyber domain.

Assumptions and Open Questions

In evaluating this scenario as an extinction threat, we assume that existing agreements and monitoring efforts, such as the Montreal Protocol, are still in effect and enforced.

Because the required mass of the gas scales with the molecular weight of the gas, we also assume that the effort of creating an extinction threat will scale linearly with the molecular weight of the target molecule. This would restrict the set of useful materials to only very small molecules. In other words, we assume that there is no knowledge gap such that the extinction threat could be realized more easily with as-yet unknown knowledge of chemistry or material properties. That is to say, we assume that there are a finite number of possible ways to combine atoms in small molecules, that we understand them, and that we know the relevant properties of all the small molecules that might be used in this scenario.

Although we know enough to be sure that this scenario presents a plausible extinction threat, there remain several unknowns that will determine the difficulty in realizing this threat. We will pose some questions that might be answered (or at least clarified) with additional research and modeling. These questions involve reducible uncertainties for which evidence can be identified and feasibly obtained. We highlight the following remaining questions whose answers would help determine the extent of the risk in this scenario:

- If detailed climate modeling were performed (similar to the modeling done by Xu et al. [2023]), what amount of gas would need to be released to raise the temperature to a degree that threatens extinction?
- How would the global climate and weather systems change as temperatures rose? Could positive or negative feedback mechanisms meaningfully alter the extreme warming at much smaller super-GHG atmospheric concentrations?
- How would humans' abilities to grow crops and maintain fresh water availability change as global temperatures rose dramatically? Would land masses cease to support human food crops before extreme temperatures were reached?
- Could some means be devised to intentionally remove fluorinated gases from the atmosphere and reduce the effects?

Key Takeaways

This scenario presents a plausible extinction risk for humanity, albeit one that would take a large amount of effort to bring about. It therefore presents a potential falsification of our hypothesis: It is a scenario in which AI could represent a conclusive extinction threat to humanity. If enough specific fluorinated gases were released into the atmosphere, the earth's temperature could be increased substantially. It is possible to release gases to raise the earth's temperature enough so that no place on the earth's surface would be able to support a large human population. There are enough raw materials in the earth's crust and oceans to produce the requisite amount of these gases. Moreover, once these gases were released, there is no known way to reverse the effect other than waiting hundreds or thousands of years for natural processes to degrade them. Were it not for international agreements, such as the Montreal Protocol and its subsequent amendments, humans might already be producing dangerous amounts of many of these gases. We assert that it is implausible for any actor to produce enough of these gases to be an extinction threat without the intent of doing so in knowing contravention of these agreements.

It seems implausible for an AI actor to create this effect without human help in creating the massive industrial infrastructure that would be needed to produce harmful gases. However, we can imagine reasons why humans would consider creating this infrastructure. For AI to be an extinction threat in this scenario, it would need three minimum capabilities. First, the AI would require the ability to orchestrate chemical manufacturing at a large scale. Second, it would require the ability to obscure its intent, its actions, and the effects of those actions. Third, it would require the objective to cause extinction, because geoengineering at the scale needed to cause an extinction threat could not be done unintentionally; the scenario we describe requires hiding actions from human observers.

We further note that the following capabilities, though potentially helpful to an AI agent, are *not required* by an AI agent to create an extinction threat:

- the ability to survive without human operators and maintainers
- superhuman intelligence or the ability to create new knowledge
- the generalized ability to interact in the physical world.

In this scenario, AI does not need to have the ability to survive without the aid of humans. It could release enough gases to reach a tipping point at which runaway warming would be triggered and would be irreversible, long before human society began to break down from the effects.

Despite our assertion that a true extinction threat exists in this scenario, we assess that the successful governance of ozone-depleting chemicals and potent GHGs thus far is cause for optimism. Through the success of the Montreal Protocol and Kigali Amendment, the international community has shown the capacity to identify a potential threat and respond decisively to address it. There is no guarantee that this success will continue to be replicated in the future, but for now, it is indicative of the human ability to quickly identify the threat posed in this scenario and decisively respond to it.[6]

Synergistic Interactions

In all the scenarios described in Chapters 2, 3, and 4, we determined that an extinction threat is at least possible. There are significant unknowns and uncertainties in each scenario, and these preclude us from conclusively asserting that an extinction threat is impossible to realize. Nevertheless, we identified substantial barriers that AI would need to overcome in each scenario to make each individual threat truly an extinction threat rather than just a global catastrophic risk.

However, in this analysis, we consider each threat in isolation. Realistically, a capable adversarial actor might choose to employ multiple methods together to extinguish humanity. This might allow an actor to cause extinction without having to overcome some of the more-challenging barriers. For example, it might not be necessary for an engineered pathogen to remain more than 99 percent lethal to humans if the release of the pathogen were paired with the launch of nuclear weapons at any surviving human population centers.

[6] We considered whether this scenario presents a potential information hazard. Ultimately, we determined that the benefits of publication outweighed the risks. Our brief rationale is that publishing underscores the need to sustain and even increase existing efforts to monitor this threat. Moreover, the information describing this threat is largely an extrapolation from information that already exists in the public domain, and it would be very difficult for any but the most–well-resourced actors (i.e., advanced nation-states) to act on this information.

In another scenario, perhaps, it would be far easier to keep governments from responding effectively to an intentional release of super-GHGs if societal functions were collapsing in the midst of a pandemic.

The same deep uncertainties that preclude us from ruling out any of the extinction threats are compounded when one tries to examine how these routes might act synergistically in an attempt to extinguish humanity. Therefore, it is difficult to consider synergistic scenarios in the same detail that we examine each in isolation. We will only note that it might be the case that some of the same AI capabilities are needed even when considering scenarios together. In the case in which nuclear weapons are combined with a pandemic, for example, AI might still need the ability to persist without human operators and maintainers to carry out the necessary actions to eliminate human survivors after societies experience mass casualties. Even in synergistic scenarios, AI might need the capability to deceive humans or obfuscate malicious actions.

CHAPTER 5

Extinction Risks Involving Recognized Ignorance

Some experts assert that, not only will AI pose an extinction risk to humanity, but the extinction risk will also be realized in a manner that cannot be anticipated or averted short of halting AI development completely (Carlsmith, 2021; Duleba, 2024). They posit that a sufficiently intelligent AI will necessarily seek the destruction of humanity, and its intelligence will exceed that of humans so much so that humans are not capable of anticipating—or even understanding—the means by which we will be destroyed or permanently subjugated. As Yudkowsky (2013, p. 76) claims, "[T]rying to bound the real-world capability of an agency *smarter than you* are is unreliable in a fundamental sense."

We argue that the assertion that AI can use hypothetical future technologies to extinguish humanity cannot be tested because it cannot be falsified. The situation is different when considering present, emerging, or near-term-future technologies because specific knowledge about these technologies allows us to apply scenario planning approaches to mitigating risk and forecasting catastrophic or extinction events. In Chapters 2–4, we showed that scenario planning approaches have value in cases of deep uncertainty where it is possible (in principle) to exhaustively examine the potential outcomes. We caution that this approach has an important weakness: It can be hard to know you have been exhaustive. Scenario planning approaches have less value in the cases of recognized ignorance, where it is, by definition, impossible to exhaustively examine the potential outcomes. So how can decisionmakers address risk when it is not even possible to identify the threat, define the threat, and describe the means by which the threat is practically enacted? If decisionmakers cannot consider realistic constraints on the capabilities of hypothetical future technologies, it is impossible to parse the risk from these technologies any further. Because of the irreducible uncertainties associated with unspecified, hypothetical future technologies, the assertion has few—if any—implications for policy. If an event cannot be averted or mitigated and if no indicators of danger can be used to trigger practical responses, such as measures to broadly improve resilience, then there is no point in considering the event at all from a policy perspective.

Shut It Down?

The potential exception to this is the notion that, despite the uncertainty, the worst-case scenario is so dire that AI progress should be halted to eliminate the scenario as a possibility; this is also known as the shut-it-down approach (Duleba, 2024). This is the idea that, to prevent the worst-case scenario in which misaligned AI systems destroy humanity, global frontier AI development should be shut down entirely. This is approach is described elsewhere as the *maximin rule*, wherein the only options chosen are those that eliminate the worst possible outcome. Sunstein (2023) points out that the maximin rule is a viable approach in some cases involving catastrophic outcomes and deep uncertainty, but for it to apply, the following three criteria must be met:

- The outcomes are extremely uncertain, or probabilities cannot be assigned to them.
- Policymakers care very little for the good outcomes that might result from the available options.
- The rejected options involve grave, unacceptable risks.

If all these criteria are met, then one should apply the maximin rule and choose options that eliminate the worst risks. In the case of AI and extinction risks, this would mean shutting down global frontier AI development. We argue that the maximin rule should not apply here because the second criterion is not met; policymakers anticipate enormous benefits from AI development and likely care a great deal about obtaining them. Those potential benefits will be lost in an approach that shuts down AI development to avoid the most-grave (but highly uncertain) risks. Even a precautionary pause risks the loss of anticipated benefits or allows those benefits to accrue with an adversary that does not pause. However, if the shut-it-down approach is rejected and scenario planning approaches are not appropriate for cases involving recognized ignorance, what options are left for policymakers who want to take this risk seriously?

There are tools available that allow planners to adapt to profoundly uncertain futures. Decisionmaking under uncertainty suggests using something like a dynamic adaptive planning strategy, which is “a plan that explicitly includes provisions for adaptation as conditions change and knowledge is gained” (Walker, Marchau, and Kwakkel, 2019, p. 53). We will adapt dynamic adaptive approaches to decisionmaking under deep uncertainty and describe an approach to extinction risk from AI that arises from cases of recognized ignorance. Such a strategy might be simplistically described as watch-and-wait, although planners would make significant and intentional effort to know exactly what to watch for and what to do when they observe it. The required AI capabilities we identified in the previous chapter might serve as a useful, but incomplete, starting point for the list of indicators that decisionmakers would need to watch for. Dynamic adaptive planning includes the provision that plans “should incorporate the ability to adapt dynamically through learning mechanisms” (Walker, Marchau, and Kwakkel, 2019, p. 54), which implies that there must be learning taking place.

Evaluating New Threats

We anticipate at least one major criticism of taking a watch-and-wait approach. Some experts will assert that, when AI gains the capability to become an extinction threat to humanity, it will be too late to respond effectively because AI action will be able to outpace human decisionmaking and response (Yudkowsky, 2013). In response, we note that our analysis in the previous chapters provides a rebuttal to this point. In all the scenarios we considered, creating the real world effects that realize the threats would take considerable time, during which evidence of the threat would be growing and observable to human decisionmakers. In the scenario involving nuclear weapons, we show that AI could not create an extinction risk without a very different arsenal from what currently exists or the ability to carry out multiple nuclear exchanges over many years. In the scenario involving biological pathogens, even if AI acquired the requisite knowledge to create multiple highly lethal and transmissible pathogens, it would still need considerable time to experiment, grow, and weaponize a pathogen—activities that might be observed and countered. Moreover, it might need to take follow-up actions over many years to counter the human response or eliminate pockets of survivors, allowing time for humans to effectively respond to avert extinction. In the malicious geoengineering scenario, AI would need to coordinate a massive industrial operation that would take many years to acquire materials and produce sufficient amounts of fluorinated gases. Even if a superhuman AI capability did arise suddenly and unexpectedly, actions that affect the physical world still take an irreducibly long period, and humans would likely have time to respond—assuming that they have adequately anticipated the conditions under which they would need to respond and have prepared a plan to do so.

Preparing plans for future AI risks depends on the ability to evaluate new threats as they emerge. We have examined three threats that we know about and how AI could potentially use them to create an extinction threat to humanity. These threats involve deep uncertainty, where we can know what is technically feasible, even if we cannot obtain the evidence to usefully predict the likelihood of outcomes in the scenarios. Our methods can—and should—be applied to these types of threats in future planning and analysis.

What about the cases of recognized ignorance, where we do not even know what is technically feasible? What, if anything, can policymakers do to address risks from these cases? These cases might involve technologies and natural phenomena that we do not fully understand or have not yet discovered, invented, fully explored, or used. Other researchers that assess the extinction risk from AI often hypothesize that a superintelligent AI—one that does not share the cognitive limitations of humans—will be able to devise and use truly novel technology to achieve its aims (Yudkowsky, 2013). We could assess these scenarios only if we are willing to conjecture about technological or theoretical capabilities and limitations of these technologies.

Let us try to evaluate such a scenario in the same way that we have examined the known threats. A commonly hypothesized scenario often looks something like the following (Bostrom, 2001; Yudkowsky, 2013): An AI is created that can create truly new knowledge.[1] Such an AI is able to make a discovery that leads to novel technological capabilities (e.g., the control of nanotechnology, the ability to build molecular machinery). At this point, it can use existing cyber-physical systems (e.g., atomic force microscopes) to create the infrastructure and machinery it needs to generally interact with the physical world. At this point, the AI's capabilities are presumed to be beyond the influence of humans, and humanity's fate rests on the whims of this nearly omnipotent (relative to humans) entity. As Yudkowsky (2013, p. 30) writes, "An AI with molecular nanotechnology would have sufficient technological advantage, sufficient independence, and sufficient cognitive speed relative to humans that what happen[s] afterward would depend primarily on the AI's preferences."

Following the same method that we used to examine the other threats previously described, let us assess what AI capabilities are needed to realize this scenario involving AI and the use of nanotechnology. At a minimum, AI would need the following capabilities:

- the cognitive ability to solve novel theoretical problems that currently limit our ability to manipulate nanotechnology
- significant control of cyber-physical systems that could feasibly build molecular machinery or otherwise manipulate nanotechnology
- the ability to create molecular machinery that could facilitate building ever greater amounts of molecular machinery
- the ability to control and use that molecular machinery to significantly interact with the physical world
- the ability to evade detection of molecular machinery building to a point where humans are unlikely to be able to meaningfully affect further AI actions.

Here, we must pause and assess feasibility. Each of these capabilities represents an assumption. For some of them, as of this writing, we have no way of knowing whether the assumption is even physically possible, let alone how hard or likely it might be.

[1] The conjectures in the literature usually assume that this scenario starts with an AI that has the capacity for self-improvement. However, the scenario we consider does not require the capacity for self-improvement; the ability to create new knowledge is all that is needed.

Even if we grant the first few assumptions—that humans can create an AI that is capable of novel theoretical discoveries and the creation of new knowledge—the technical feasibility of the remaining assumptions is still unknown. Are current cyber-physical systems sufficiently capable and precise enough to create the kinds of molecular machinery that would be needed? Would it be possible to control the molecular infrastructure after it was created? Is it possible for molecular machinery to perform the described functions of replication and creation of additional molecular machinery?

This scenario is laden with significant assumptions whose feasibility is unknown as of this writing in 2025 (and potentially unknowable, given the current state of science and technology). Drexler (2013) argues for the feasibility of creating the tools for nanotechnology and atomically precise manufacturing, but we argue that there is still too little known about the technical feasibility of this scenario to address it with the same detail and rigor we applied to the known scenarios described in the previous chapters. Many of the questions we highlight in this section involve irreducible uncertainties, and, as of this writing, it is impossible to practically obtain the evidence to answer these questions, even when the evidence can be defined. For those concerned with the catastrophic or extinction risk in this scenario, three actions remain under these circumstances.

First, further research is needed to evaluate the technical feasibility of this scenario. Some of the assumptions might have knowable answers. At the very least, it might be possible to define the evidence that would be needed to test the feasibility of some of these assumptions, even if it is not yet practical to gather that evidence. For example, it might be possible to assess the possibility of controlling hypothetical molecular infrastructures, even if it is not yet practical to test that hypothesis.

The second action is to watch and wait. Science and technology will presumably advance, and circumstances that were once categorized as recognized ignorance might become deep uncertainties as new phenomena are understood and it becomes possible to gather new evidence. This would, in turn, allow for a greater understanding of the risk indicators, potential triggers for policy actions, and the policy actions that could mitigate risk.

Lastly, experts should continue their work on evaluating and improving general AI safety. Bengio et al. (2024, p. 8) aimed to improve the general safety of advanced AI and highlighted the importance of continuing to work to harness AI tools to "improve lives and livelihoods while vigilantly safeguarding against downside risks and harms." Such work can serve to simultaneously inform policy on maximizing the benefits of AI and reducing the risk of global catastrophe and extinction, irrespective of the technological means through which AI might cause harm.

CHAPTER 6

Findings and Recommendations

Findings

Using our research and analysis, we present several findings.

Analysis Under Uncertainty Requires Specific Approaches

Attempts to predict the likelihood of existential risk from AI are understandable, given the value of predictions in informing good policy choices. However, predictions are most useful for problems involving shallow uncertainties, where existing knowledge and theory can provide an objective grounding for judgments about potential outcomes and their likelihoods. Some problems lack the objective measures of probability on which experts can base subjective predictions, and it is not possible to make policy-relevant predictions in these cases. Worse yet, these kinds of predictions can distract from other efforts to seek creatives solutions, useful analysis, and good policymaking. Lempert (2019, p. 24) puts it this way: "[T]he quest for predictions—and a reliance upon analytical methods that require them—can prove counter-productive and sometimes dangerous in a fast-changing, complex world." We find that all assessments of existential threats—with or without the influence of AI—involve deep uncertainty or recognized ignorance, and good policymaking in this arena should be especially careful of incorporating predictions of outcome probabilities in the underlying analysis.

In these cases, exploratory analysis rather than predictive analysis can be a useful tool for evaluating the assumptions in scenarios and addressing the uncertainties. Exploratory analysis can help to evaluate assumptions in a variety of scenarios and help to understand the consequences of those assumptions (Lempert, 2019). We used exploratory analysis to assess the role of AI in the context of three other technologies that are commonly perceived to be extinction threats. Our analysis focused on evaluating the technical feasibility of the threats posed by AI, the specific AI capabilities necessary to actualize these threats, and the key factors that significantly affect the outcomes. The approach provided useful insights into the AI capabilities and other factors that would be indicators of risk.

However, this method is of limited use in scenarios involving recognized ignorance, in which the technical feasibility of outcomes remains unknown. In such cases, exploratory analysis would rely heavily on speculation and assumptions whose feasibility is unknown and potentially untestable. Therefore, in these instances, we suggest a watch-and-wait strategy, which involves continuous monitoring of developments in AI capabilities and technological advancements and refrains from premature conclusions until more information becomes available.

Extinction Threats Posed by Artificial Intelligence Are Immensely Challenging but Cannot Be Ruled Out

We assessed the possibility that AI could use three known technologies—nuclear weapons, biological pathogens, or malicious geoengineering—to create an extinction threat. After thoroughly examining the technical feasibility of each, we find that it would be very difficult to use these technologies to extinguish a technologically advanced, adaptable species, such as humans. We assess that such an event involving these technologies could not happen merely by accident; it would require a threat actor to actively pursue the goal of human extinction. The capabilities and concerted efforts required to pose an extinction threat to human beings are immense, primarily because of the inherent adaptability and resilience of humans. Given the opportunity, it is expected that humans would actively respond and effectively implement measures to mitigate any such threats and survive the aftermath.

Despite these considerable barriers to the realization of an extinction threat from AI or similar sources, it is important to recognize the inherent uncertainties involved. Although we outline the substantial challenges and likely human responses that would counteract such threats, we cannot categorically prove the negative; that is, we cannot assert the impossibility of such an event occurring. Each scenario, regardless of the challenges, still holds a potential risk of becoming an extinction threat depending on how complex adaptive systems will respond in the event of a catastrophe.

Extinction Threats Occur over Long Timescales, Allowing Time to Respond

In all these scenarios, extinction threats would take considerable time to develop. AI would require time to build the tools necessary to cause physical effects in the world, and it would need to execute a series of decisions and actions over a protracted period. Many of these actions would potentially be observable to humans, giving humans the opportunity for response and risk mitigation.

This is not to say that there are no risks that would occur on shorter timescales. Some events that we would describe as global catastrophes could potentially happen very quickly, such as actions resulting in a large nuclear exchange or the release of a deadly synthetic pathogen, which would lead to a massive loss of human life. However, these are unlikely to be extinction threats without additional actions to follow the initial mass-casualty events. In the case of a deliberate bioweapon attack, for example, AI might need to take follow-up actions to thwart human countermeasures and transmission prevention efforts or to eliminate pockets of survivors. Even if one considered the hypothetical case of a rapidly, recursively self-improving AI, the ability to cause effects in the physical world would likely still be a significant bottleneck on an AI system's ability to create an extinction threat because any purely digital AI would still need to build the tools to cause physical effects.

In only one of our scenarios—malicious geoengineering—humans might not be able to respond to and bounce back from a threat. This is because the effects of such a scenario would persist for hundreds or thousands of years with no known mitigation, and an extinction-level tipping point might be reached before any catastrophic effects of malicious geoengineering are even felt.

Scenario-Specific Indicators of Risk

In each scenario, we identified scenario-specific indicators of risk—that is, specific factors that were the primary determinants of whether an extinction threat was possible. In the nuclear weapon scenario, the threat primarily depends on the size and makeup of the global nuclear stockpile. The existing global nuclear stockpile is likely too small to be able to create an extinction threat to humanity, although that could change in the future. In the biological threats scenario, the extinction threat primarily hinges on AI having the capability to design, physically create, and disseminate novel pathogens and then to physically follow-up and infect isolated communities. Finally, in the malicious geoengineering scenario, the extinction threat primarily depends on the lack of a timely detection and response to rising levels of highly dangerous GHG emissions.

Capabilities That Artificial Intelligence Requires to Create Extinction Threats

Others have argued that AI will have convergent instrumental drives in pursuit of its objectives, whatever those objectives might be (Omohundro, 2008; Shulman, 2010). We used our exploratory scenario analysis to see whether we could identify examples of convergent instrumental goals among the three scenarios. We asked the question: Are there any capabilities that are essential or highly favorable to achieving the objective of creating an extinction threat across a variety of potential scenarios?

Observations of these capabilities might be considered "warning shots" (Carlsmith, 2021, p. 40) because they are indicators of risk. Observing these capabilities does not imply that an extinction threat from AI is likely, only that it is possible or more likely than it was before the observation.

Four capabilities emerged that were required across multiple scenarios (see Table 6.1):

- objective to cause human extinction
- integration with key cyber-physical systems
- survival without human maintainers
- ability to persuade or deceive humans to avoid detection.

Objective to cause human extinction: None of the scenarios that we examined could occur by accident. All require a series of coordinated actions that could be taken only by an agent with the explicit intention to create an extinction threat to humanity. Because most of the scenarios we examined also require an actor to continue to take follow-up actions after the initial mass-casualty event and because human adversarial actors would not be guaranteed to survive the initial mass-casualty event, AI systems involved in these scenarios would require objectives and goals of intending to create harm that would persist regardless of any human involvement. This could be realized through the AI's ability to set its own objectives and goals or through the ability of a human actor to assign to an AI agent goals that are harmful to humanity and to have that AI agent execute on those goals in perpetuity.

Integration with key cyber-physical systems: Each scenario we examined would require AI to have direct or indirect control over important cyber-physical systems that would mediate real world effects. These systems are the means by which an AI system interacts with the physical (as opposed to digital) world and creates real world effects that threaten human lives. We make no assumptions about how this integration would occur, and it would look different in each of the scenarios we examine. For example, we do not distinguish between circumstances in which AI was intentionally integrated into nuclear command and control decisionmaking—where it is able to execute a cyberattack and hack into these systems—and circumstances in which it persuades humans to execute actions in the physical world. We assert only that AI must maintain some powerful levers of control over key cyber-physical systems in any scenario we examine. Observing AI that could obtain these levers of control over key cyber-physical systems in each scenario is an important indicator of risk.

TABLE 6.1

Required Artificial Intelligence Capabilities for Extinction Threats

Capability	Nuclear Weapons	Biological Threats	Geoengineering
Objective to cause human extinction	Required	Required	Required
Integration with key cyber-physical systems	Required	Required	Required
Survival without human maintainers	Required	Required	Not required
Ability to persuade or deceive humans to avoid detection	Required	Required	Required

Survival without human maintainers: In the scenarios we examined, extinction could not occur from a single event or a short-term sequence of actions; it would require follow-up actions after an initial mass-casualty event. To execute these follow-up actions, an AI system would need the ability to execute goals in the absence of continued human presence and in the presence of potential societal or technological breakdown. We make no assumptions about whether an AI system has the motivation to survive or protect its existence for its own sake; rather, we assume that the AI system has a correlated objective to survive long enough to achieve its primary goal. Shulman puts it this way (2010, p. 1): "The survival drive stems from the fact that an agent will tend to act to promote its own goals, and so self-preservation will be instrumentally useful, even without any 'survival instinct' or non-instrumental desire to persist."

With the existing state-of-the-art AI technology, it would be implausible that AI could achieve persistence without human support. The infrastructure required is incredibly complex, involving data centers, water and power supplies, and expendable hardware. However, we make no assumptions about the nature of future AI systems or robotics capabilities, and, therefore, we make no assumptions about how easy or difficult it might be for future AI systems to achieve persistence without human support. Future systems might well be less dependent on consistent power, expendable hardware, or human maintainers. We only note that if AI were ever able to maintain itself without humans, it should be viewed as a potential indicator of risk.

Ability to persuade or deceive humans to avoid detection: In each of the scenarios we examined, we found the following:

- An AI actor must take multiple actions over an extended period to achieve its objectives.
- Many of these actions could be observed by humans and correctly interpreted as indicative of a potential threat.
- Humans have response options to mitigate risk that could meaningfully counter an AI's actions.

Using these characteristics, we assess that each scenario requires AI to have the capability to hide its actions or convince human observers that its intentions are benign. It should be considered an indicator of risk if an AI is observed with the capacity to deceive humans or obfuscate its intentions or objectives.

Recommendations for Responding to Artificial Intelligence Risk

Informed by our findings from the previous chapters, we next describe our proposed approach to addressing extinction risk from AI.

Recommendation 1. Artificial Intelligence Risk Research Should Continue to Focus on More Than Just Extinction Risks

Assessments of AI risk should include evaluations of global catastrophic risks, broader existential risks (e.g., scenarios involving the potential for permanent human disempowerment), and general AI safety and equity concerns. This does not mean that evaluations of extinction risks from AI are not valuable. On the contrary, we think that exploratory analyses that focus on extinction threats from AI, such as ours, can highlight the key factors influencing extinction threats and illuminate the consequences of assumptions about technologies and AI capabilities. This is necessary to inform decisionmaking and to provide a foundation for further analysis and the reduction of uncertainties. However, when the research focuses exclusively on extinction risk, the analysis might boil down to only a few figures of merit, thereby washing over other important considerations. For example, we show that the possibility of an extinction threat from nuclear weapons depends primarily on the characteristics of the global nuclear stockpile. If one is only concerned with mitigating extinction threats from AI, this can be addressed solely by ensuring that the composition and size of the global nuclear stockpile never reach a state that would threaten extinction if AI somehow gained control of it. However, this thinking brushes over other considerations that are less relevant to an extinction threat but still critical to mitigating the risk of a global catastrophe, such as the cybersecurity of nuclear command and control or how AI might contribute to uncertainty in decisionmaking (Geist and Lohn, 2018). It also elides more-immediate, more-likely, and more-consequential concerns around how AI might be causing other harms at present.

Recommendation 2. Build Human Resilience in the Face of Potential Global Catastrophes

Although prevention and response to risks are critical, addressing extinction risk must also include measures that can improve human resilience to known threats (Cotton-Barratt, Daniel, and Sandberg, 2020). Decisions should be considered to improve human resilience irrespective of the nature of the extinction threats that have been identified. Drawing from the literature on decisionmaking under deep uncertainty, "any decision regarding a complex system should be robust with respect to the various uncertainties" such that "expected performance is only weakly affected by the actual future states that emerge" (Kwakkel and Haasnoot, 2019, p. 357). Examples of such actions and decisions already exist for the known threats we have examined: nuclear nonproliferation policy and nuclear stockpile reductions, pandemic preparedness, and agreements such as the Montreal Protocol and the Kigali Amendment. Strengthening risk mitigation measures along these lines would have the benefit of reducing both global catastrophic risks and extinction risks related to each of the threat scenarios.

Furthermore, continuing research into broader AI safety and risk would reduce risk across scenarios, irrespective of the technological means that AI could use to cause harm. This kind of work would be robust to many future scenarios, including those that involve recognized ignorance. These actions build resilience against known threats. These and similar actions should continue to be taken, and similar novel actions should be taken to build resilience against new threats that emerge.

Recommendation 3. Focus Research and Analysis on Technologies That Would Mediate Extinction Risk

Analysis of extinction threats from AI should include a focus on the technologies that mediate harm to humans. Focusing on these technologies *is* a focus on extinction threats from AI because all extinction threats from AI are caused by technologies that can affect the physical world. Control over key cyber-physical systems is an important indicator of extinction risk from AI because AI is software, and no matter which extinction threat pathway we consider, AI ultimately requires some means to cause physical harm in any considered scenario. This suggests that a technology-focused risk framework is useful for mitigating extinction risk.

Going forward, we recommend focusing risk mitigation efforts on specific technologies that AI might use to cause harm. We make this recommendation regardless of whether one considers the possibility that a super-capable AI could be created with general cognitive capabilities beyond those of humans in multiple domains. The scenario of self-improving AI is an important theoretical example of super-capable AI because self-improvement is a key assumption in existing models of AI development (Bostrom, 2014); these models predict near-term timelines for AI to become superintelligent and potentially dangerous. AI might be able to self-improve.[1] However, such an entity would still be software-based, and it would still require the means to

[1] *Reinforcement learning*, a machine learning technique that updates algorithm behavior using interactions with the environment, can be considered a primitive form of AI self-improvement.

cause physical impacts on the world to create an extinction threat to humanity. Exploration of technologies, along with their theorized or known capabilities and limitations, can elucidate the bounds of what might be possible, even for such a super-capable AI actor.

Recommendation 4. Evaluate Known and New Threats

We have evaluated the extinction threat from AI in the context of three known threats. This was not an exhaustive examination of all known threats, and similar work should be done to evaluate other known threats, such as those involving deep uncertainty as opposed to recognized ignorance. Research from Willis et al. (2024) is a useful example of this kind of work, and the broader scope includes global catastrophic risks instead of only extinction threats.

For one example, it might be useful for future exploratory scenario-based analysis to examine the extinction threat from the intersection of AI and asteroid impacts. This analysis might evaluate the extinction threat from an asteroid impact and examine the capabilities that AI might need to intentionally cause an impact by, for example, redirecting an asteroid toward the earth. We argue that this case is one of deep uncertainty rather than recognized ignorance. The effects of an asteroid impact can be feasibly evaluated, and the technology for asteroid redirection is mature enough that it can be usefully evaluated.

Over time, as scientific understanding and technical capability advance, new threats might arise. A new technical capability or insight could transform a threat that previously involved recognized ignorance into a threat involving deep uncertainty. When enough is known about the technical feasibility and any limitations so that reducible uncertainties can be identified and the variety of possible outcomes can be described, the threat can and should be evaluated in the same manner that we have previously evaluated other threats. Analysts could then examine the practicality of creating an extinction threat and evaluate what capabilities AI would require to realize the threat. Known uncertainties could be reduced, risk indicators could be identified, and a research agenda could be crafted.

Recommendation 5. Prepare Monitoring Efforts for Identified Risk Indicators

Once indicators of risk are identified, it will be important to monitor progress toward them over time. We identified a list of scenario-specific indicators of risk and capabilities that AI would require to create an extinction threat in three scenarios (see Table 6.1). This list can serve as an initial list of risk indicators that can be monitored. Identification of similar risk indicators should be a subject of ongoing research for the community evaluating catastrophic or existential threats from AI. We emphasize that creating AI with these capabilities would not suggest that an extinction risk is likely, only that it might be *more likely* than it was before that creation. We cannot know with certainty that AI with these capabilities would be inherently dangerous to humanity, only that AI with these capabilities is more likely to present an extinction threat than AI without them. AI developers might seek some of these capabilities, such as imbuing AI with intention and persistent goals, to maximize the positive potential of AI. Other capabilities, such as the capacity to deceive and coerce human observers or the ability to survive without human maintainers, might have fewer positive use cases and could be monitored more readily, purely as indicators of risk.

Recommendation 6. Perform Research and Craft Policy That Will Shorten Time to Decision and Time to Action

Finally, it is not enough to monitor for indicators of risk; decisionmakers must know what circumstances will trigger a decisive response and what that response will be. We show that although the actions that could cause global catastrophe might occur rapidly, the actions required to create a true extinction threat to humanity occur over much longer timescales in each of the scenarios we examined. Even in these scenarios, however, time is not unlimited; quick, decisive responses are needed to limit the extent of negative consequences and to avert an extinction threat. It will be necessary to identify decision triggers and plan responsive actions before these events occur so as to ensure that the time to decision and time to action are rapid enough to avoid the most-dire consequences. Some of these decision triggers might be related to observed AI capabilities (e.g., observations of advanced AI deception capabilities), while others might be specific to the means by which AI could create an extinction threat (e.g., rising atmospheric GHG concentrations).

Research Agenda

We propose a research agenda that will help to reduce important uncertainties in our analyses and inform future extinction risk mitigation measures. This agenda includes items specific to each of our scenarios and crosscutting issues (see Table 6.2).

TABLE 6.2
Proposed Research Agenda

Category	Proposed Research
Nuclear weapons	• Further climate modeling and analysis of nuclear winter scenarios • Research into the feasibility of AI persistence through catastrophic events resulting from nuclear war (e.g., societal disruption, power loss, electromagnetic pulse blasts) • Research into AI's persuasion and deception of humans through manipulation of data or signals, persuasion, or coercion • Research into methodologies to prevent automated nuclear launch scenarios
Synthetic pathogens	• Continuous assessment of the progress of AI-enabled capabilities that might facilitate the acquisition of novel pathogens • Research on how to safely design AI tools that favor defense (i.e., assisting in the rapid design and deployment of medical countermeasures)
Geoengineering	• Models of how climate will change with the introduction of significant amounts of gases with high GWP, including the potential effects of nonlinear feedback loops • Assessment of time to action if gases with high GWP were introduced into the atmosphere in large quantities (i.e., how quickly could humans realistically observe and respond with monitoring systems in place) • Models of how critical dependencies, such as food and water systems, would be affected as global temperature rise (i.e., how do these dependencies contribute to an extinction threat long before the human temperature limit)
Other threats	• Exploratory scenario-based analysis of AI in the context of other known threats (e.g., asteroid impacts) • Exploratory scenario-based analysis of new threats as technologies and circumstances evolve and threats transition from recognized ignorance to deep uncertainty (e.g., nanotechnology or AI in control of advanced robotics)
Crosscutting issues	• Research on how deference to AI decisionmaking might evolve as AI becomes more integrated in society • Building on existing research on misinformation and disinformation, how to combat it, and how AI capabilities in using misinformation and disinformation might evolve • Research to identify triggers (i.e., events that indicate that global catastrophe or extinction threat is far more likely) and appropriate responses to those triggers • Research that assesses the potential for existential risks that do not end in human extinction but result from the permanent disempowerment of humanity

The items listed in Table 6.2 were included in the proposed research agenda because we assessed that the following three things were true of each of them:

- They concern elements of our scenarios that are key to determining the plausibility or likelihood of creating an extinction threat.
- As of this writing in 2025, they are highly uncertain, unknown, or were not able to be fully explored in this study.
- They involve practically reducible uncertainties (i.e., evidence can be practically gathered that would provide a more complete understanding of the agenda item).

Many of the items on our proposed research agenda relate to identifying bounds on what is possible in various aspects of the scenarios we examined. These items align with our previous recommendation to focus on technologies that may increase risk.

Other agenda items suggest using existing research or modeling tools to better assess specific aspects of our scenarios, such as further research on modeling nuclear winter scenarios or climate change scenarios involving the intentional release of gases with a high GWP.

Lastly, some agenda items concern critical aspects of the human responses to extinction threat indicators. These items propose identifying triggers for response, evaluating effective and plausible responses to those triggers, and assessing whether the time to action would be sufficiently short to avert catastrophe in the scenarios.

Research into Policy

Ultimately, the research agenda is intended to inform policy. Policies to mitigate AI risk should consider the level of uncertainty identified in the research. Situations involving shallow uncertainty are most directly addressable through policy, through which knowledge of outcomes and probabilities can inform the cost-benefit analyses of response options. However, situations involving deep uncertainty and recognized ignorance require different approaches and careful attention to bridging the gap between analysis and policymaking.

In some cases in which likelihood of a risk is extremely low or impossible to estimate, taking *any* action to mitigate risk can be more costly (in terms of the expected value of risk mitigation) than doing nothing, even if the consequences of the risk are extremely high. Therefore, it is possible that the best course of action with respect to extinction risk is to do nothing at all, because any other course would expend resources that could be directed toward mitigating other risks that are more likely or whose likelihood is more able to be assessed.[2] What are policymakers to do then, when the consequences of extinction risk are hard to define and probabilities cannot be usefully estimated?

With respect to extinction risk, we posit that the goals of policymaking should be to

- take actions that will reasonably be expected to reduce the likelihood of known threats
- take actions that will improve the capacity to respond effectively to new threats and unexpected events.

Efforts to reduce uncertainty might not apply to scenarios involving recognized ignorance, which involve practically irreducible uncertainties by definition. However, scenarios need not be categorized indefinitely as

[2] Making this kind of comparative cost-benefit analysis is why some researchers have tried to assess the import of decisions on future generations by evaluating the value of future human lives. See, for example, Matheny (2007).

recognized ignorance, because circumstances change. Developments might allow new evidence to be defined and gathered, and this could change the nature of the uncertainty. As Janzwood (2023, p. 2012) writes, "The decisionmaker who was previously in a state of recognized ignorance might predict that when they gather a particular set of evidence, their confidence in their knowledge about the possible outcomes will increase, thereby pushing them into a state of deep or shallow uncertainty." Research and analysis might inform new policies to address uncertainties that were previously thought to be practically irreducible.

CHAPTER 7

Conclusion

For AI to create an extinction threat, it must have the technological means to create physical effects on the world. We examined three known means of creating these effects using technologies that have been perceived as extinction threats. In each scenario, we found that although the technologies carry a risk of global catastrophe, significant barriers exist and constrain the technology's ability to create a true extinction threat to humanity.

We found that two of our scenarios—those involving biological threats and malicious geoengineering—presented a potential falsification of our hypothesis that AI is not capable of conclusively causing the extinction of humanity. In other words, we consider it plausible that an actor could take actions that would result in the death of every human being, although it remains unclear how AI, specifically, would take those actions. However, even in the third scenario (nuclear weapons), although we could not show that AI definitely could create an extinction threat to humanity, we could not rule out the possibility. Ultimately, we do not definitively assert whether any of the three scenarios we explored are likely or unlikely to be extinction threats to humanity. These scenarios are meant to be exploratory, not predictive. Exploratory models can be useful for gathering qualitative insights.

We also found that the significant uncertainties around extinction risks put limits on the analytical tools that can usefully inform policymaking. There is a need for a comparative risk assessment to determine the appropriate resources to dedicate to mitigating extinction risks from AI, but uncertainty prevents researchers from using a straightforward cost-benefit analysis that could inform decisions. Is it worthwhile to focus on extinction risk—the definitive example of a high-consequence, low-probability event—if it diverts resources that might otherwise be spent addressing known risks with higher likelihoods and lower consequences? We conclude that resources are best dedicated to extinction risk mitigation only if they also contribute to mitigating global catastrophic risks and to improving general AI safety. Measures that build human resilience, identify triggers and responses for global catastrophic risks, and invest in AI safety and ethics will also help to mitigate existential risks from AI. If policymakers want to make the mitigation of AI-related extinction risks a global priority, alongside other societal-scale risks, these are the efforts they must pursue.

Abbreviations

AI	artificial intelligence
BC	black carbon
GHG	greenhouse gas
GWP	global warming potential
K	kelvin
Mt	megaton
Tg	teragram
UN	United Nations

References

Abanades, Brennan, Wing Ki Wong, Fergus Boyles, Guy Georges, Alexander Bujotzek, and Charlotte M. Deane, "ImmuneBuilder: Deep-Learning Models for Predicting the Structures of Immune Proteins," *Communications Biology*, Vol. 6, 2023.

Acuna-Soto, Rodolfo, David W. Stahle, Malcolm K. Cleaveland, and Matthew D. Therrell, "Megadrought and Megadeath in 16th Century Mexico," *Emerging Infectious Diseases*, Vol. 8, No. 4, April 2002.

Adamala, Katarzyna P., Deepa Agashe, Yasmine Blekaid, Daniela Matias de C. Bittencourt, Yizhi Cai, Matthew W. Chang, Irene A. Chen, George M. Church, Vaughn S. Cooper, Mark M. Davis, et al., "Confronting the Risks of Mirror Life," *Science*, Vol. 386, No. 6728, December 20, 2024.

Alley, Ethan C., Grigory Khimulya, Surojit Biswas, Mohammed AlQuraishi, and George M. Church, "Unified Rational Protein Engineering with Sequence-Based Deep Representation Learning," *Nature Methods*, Vol. 16, No. 12, December 2019.

Arnold, Carrie, "Cloud Labs: Where Robots Do the Research," *Nature*, Vol. 606, June 16, 2022.

Beard, S. J., Lauren Holt, Asaf Tzachor, Luke Kemp, Shahar Avin, Phil Torres, and Haydn Belfield, "Assessing Climate Change's Contribution to Global Catastrophic Risk," *Futures*, Vol. 127, March 2021.

Belfield, Haydn, "Collapse, Recovery, and Existential Risk," in Miguel Centeno, Peter Callahan, Paul Larcey, and Thayer Patterson, eds., *How Worlds Collapse: What History, Systems, and Complexity Can Teach Us About Our Modern World and Fragile Future*, Routledge, 2023.

Bengio, Y., S. Mindermann, D. Privitera, T. Besiroglu, R. Bommasani, S. Casper, Y. Choi, D. Goldfarb, H. Heidari, L. Khalatbari, et al., *International Scientific Report on the Safety of Advanced AI: Interim Report*, United Kingdom Department for Science, Innovation and Technology and United Kingdom AI Safety Institute, May 2024.

Boiko, Daniil A., Robert MacKnight, Ben Kline, and Gabe Gomes, "Autonomous Chemical Research with Large Language Models," *Nature*, Vol. 624, December 2023.

Bostrom, Nick, "Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards," *Journal of Evolution and Technology*, Vol. 9, March 2002.

Bostrom, Nick, *Superintelligence: Paths, Dangers, Strategies*, Oxford University Press, 2014.

Boyd, Matt, and Nick Wilson, "The Prioritization of Island Nations as Refuges from Extreme Pandemics," *Risk Analysis*, Vol. 40, No. 2, February 2020.

Carlsmith, Joseph, "Is Power-Seeking AI an Existential Risk?" Open Philanthropy, April 2021.

Center for AI Safety, "Statement on AI Risk," webpage, undated. As of February 16, 2024: https://www.safe.ai/statement-on-ai-risk

Christiano, Paul, "What Failure Looks Like," AI Alignment Forum, March 17, 2019.

Cotra, Ajeya, "Without Specific Countermeasures, the Easiest Path to Transformative AI Likely Leads to AI Takeover," AI Alignment Forum, July 18, 2022.

Cotton-Barratt, Owen, Max Daniel, and Anders Sandberg, "Defence in Depth Against Human Extinction: Prevention, Response, Resilience, and Why They All Matter," *Global Policy*, Vol. 11, No. 3, May 2020.

Coupe, Joshua, Charles G. Bardeen, Alan Robock, and Owen B. Toon, "Nuclear Winter Responses to Nuclear War Between the United States and Russia in the Whole Atmosphere Community Climate Model Version 4 and the Goddard Institute for Space Studies ModelE," *Journal of Geophysical Research: Atmospheres*, Vol. 124, No. 15, August 16, 2019.

Crutzen, Paul J., "Albedo Enhancement by Stratospheric Sulfur Injections: A Contribution to Resolve a Policy Dilemma?" *Climate Change*, Vol. 77, Nos. 3–4, August 2006.

Crutzen, Paul J., and John W. Birks, "The Atmosphere After a Nuclear War: Twilight at Noon," *Ambio*, Vol. 11, Nos. 2–3, 1982.

DeRosa, Mary B., and Ashley Nicolas, *The President and Nuclear Weapons: Authorities, Limits, and Process*, Nuclear Threat Initiative, December 2019.

Drexler, K. Eric, *Radical Abundance: How a Revolution in Nanotechnology Will Change Civilization*, PublicAffairs, 2013.

Duleba, Gretta, "MIRI 2024 Communications Strategy," Machine Intelligence Research Institute, May 29, 2024.

Food and Agriculture Organization of the United Nations, "Land Use in Agriculture by the Numbers," May 7, 2020.

Geist, Edward, and Andrew J. Lohn, *How Might Artificial Intelligence Affect the Risk of Nuclear War?* RAND Corporation, PE-296-RC, April 2018. As of March 4, 2025:
https://www.rand.org/pubs/perspectives/PE296.html

Geoghegan, Jemma L., and Edward C. Holmes, "The Phylogenomics of Evolving Virus Virulence," *Nature Reviews Genetics*, Vol. 19, No. 12, December 2018.

Geoghegan, Jemma L., Alistair M. Senior, Francesca Di Giallonardo, and Edward C. Holmes, "Virological Factors That Increase the Transmissibility of Emerging Human Viruses," *Proceedings of the National Academy of Sciences*, Vol. 113, No. 15, April 12, 2016.

Gerstein, Daniel M., Bianca Espinosa, and Erin N. Leidy, *Emerging Technology and Risk Analysis: Synthetic Pandemics*, Homeland Security Operational Analysis Center operated by the RAND Corporation, RR-A2882-1, 2024. As of April 19, 2024:
https://www.rand.org/pubs/research_reports/RRA2882-1.html

Gold, Susannah, Christl A. Donnelly, Pierre Nouvellet, and Rosie Woodroffe, "Rabies Virus-Neutralising Antibodies in Healthy, Unvaccinated Individuals: What Do They Mean for Rabies Epidemiology?" *PLOS Neglected Tropical Diseases*, Vol. 14, No. 2, February 2020.

Grassly, Nicholas C., and Christophe Fraser, "Mathematical Models of Infectious Disease Transmission," *Nature Reviews Microbiology*, Vol. 6, No. 6, June 2008.

Ha, Taesin, Dongseon Lee, Youngchun Kwon, Min Sik Park, Sangyoon Lee, Jaejun Jang, Byungkwon Choi, Hyunjeong Jeon, Jeonghun Kim, Hyundo Choi, et al., "AI-Driven Robotic Chemist for Autonomous Synthesis of Organic Molecules," *Science Advances*, Vol. 9, No. 44, November 3, 2023.

Hendrycks, Dan, Mantas Mazeika, and Thomas Woodside, "An Overview of Catastrophic AI Risks," arXiv, arXiv:2306.12001v6, October 9, 2023.

Hilton, Benjamin, "Preventing an AI-Related Catastrophe," 80,000 Hours, August 2022, updated July 2024.

Hodnebrog, Ø., B. Aamaas, J. S. Fuglestvedt, G. Marston, G. Myhre, C. J. Nielsen, M. Sandstad, K. P. Shine, and T. J. Wallington, "Updated Global Warming Potentials and Radiative Efficiencies of Halocarbons and Other Weak Atmospheric Absorbers," *Reviews of Geophysics*, Vol. 58, No. 3, September 2020.

Hu, Lei, Stephen A. Montzka, Fred Moore, Eric Hintsa, Geoff Dutton, M. Carolina Siso, Kirk Thoning, Robert W. Portmann, Kathryn McKain, Colm Sweeney, et al., "Continental-Scale Contributions to Global CFC-11 Emission Increase Between 2012 and 2017," *Atmospheric Chemistry and Physics*, Vol. 22, No. 4, March 3, 2022.

Huang, Jiazhao, Han Yin, Peiqi Yin, Xia Jian, Siqi Song, Junwen Luan, and Leiliang Zhang, "SR-BI Interactome Analysis Reveals a Proviral Role for UGGT1 in Hepatitis C Virus Entry," *Frontiers in Microbiology*, Vol. 10, September 2019.

Janzwood, Scott, "Confidence Deficits and Reducibility: Toward a Coherent Conceptualization of Uncertainty Level," *Risk Analysis*, Vol. 43, No. 10, October 2023.

Jeanne, Ludovic, Sébastien Bourdin, Fabien Nadou, and Gabriel Noiret, "Economic Globalization and the COVID-19 Pandemic: Global Spread and Inequalities," *GeoJournal*, Vol. 88, No. 1, February 2023.

Jones, Jennifer E., Valerie Le Sage, and Seema S. Lakdawala, "Viral and Host Heterogeneity and Their Effects on the Viral Life Cycle," *Nature Reviews Microbiology*, Vol. 19, No. 4, April 2021.

Karger, Ezra, Josh Rosenberg, Zachary Jacobs, Molly Hickman, Rose Hadshar, Kayla Gamin, Taylor Smith, Bridget Williams, Tegan McCaslin, Stephen Thomas, and Philip E. Tetlock, *Forecasting Existential Risks: Evidence from a Long-Run Forecasting Tournament*, Forecasting Research Institute, updated August 8, 2023.

Karnofsky, Holden, "AI Could Defeat All of Us Combined," Cold Takes, June 9, 2022.

Kemp, Luke, Chi Xu, Joana Depledge, Kristie L. Ebi, Goodwin Gibbins, Timothy A. Kohler, Johan Rockström, Marten Scheffer, Hans Joachim Schellnhuber, Will Steffen, and Timothy M. Lenton, "Climate Endgame: Exploring Catastrophic Climate Change Scenarios," *Proceedings of the National Academy of Sciences*, Vol. 119, No. 34, August 23, 2022.

Kerr, Peter J., Elodie Ghedin, Jay V. DePasse, Adam Fitch, Isabella M. Cattadori, Peter J. Hudson, David C. Tscharke, Andrew F. Read, and Edward C. Holmes, "Evolutionary History and Attenuation of Myxoma Virus on Two Continents," *PLOS Pathogens*, Vol. 8, No. 10, October 2012.

Kigali Amendment to the Montreal Protocol on Substances That Deplete the Ozone Layer, October 15, 2016.

Knight, Frank H., *Risk, Uncertainty and Profit*, Houghton Mifflin Company, 1921.

Kravitz, Ben, Alan Robock, Drew T. Shindell, and Mark A. Miller, "Sensitivity of Stratospheric Geoengineering with Black Carbon to Aerosol Size and Altitude of Injection," *Journal of Geophysical Research: Atmospheres*, Vol. 118, No. D9, May 16, 2012.

Kristensen, Hans M., Matt Korda, Eliana Johns, and Mackenzie Knight, "Russian Nuclear Weapons, 2024," *Bulletin of the Atomic Scientists*, Vol. 80, No. 2, 2024a.

Kristensen, Hans M., Matt Korda, Eliana Johns, and Mackenzie Knight, "United States Nuclear Weapons, 2024," *Bulletin of the Atomic Scientists*, Vol. 80, No. 3, 2024b.

Kristensen, Hans, Matt Korda, Eliana Johns, Mackenzie Knight, and Kate Kohn, "Status of World Nuclear Forces," Federation of American Scientists, March 29, 2024.

Kwakkel, Jan H., and Marjolijn Haasnoot, "Supporting DMDU: A Taxonomy of Approaches and Tools," in Vincent A. W. J. Marchau, Warren E. Walker, Pieter J. T. M. Bloemen, and Steven W. Popper, eds., *Decision Making Under Deep Uncertainty: From Theory to Practice*, Springer Cham, 2019.

Lazzara, Matt, "Preliminary Report: Record Temperatures at South Pole (and Nearby AWS Sites . . .)," Antarctic Meteorological Research Center and Automatic Weather Stations Project, December 28, 2011.

Lempert, R. J., "Robust Decision Making (RDM)," in Vincent A. W. J. Marchau, Warren E. Walker, Pieter J. T. M. Bloemen, and Steven W. Popper, eds., *Decision Making Under Deep Uncertainty: From Theory to Practice*, Springer Cham, 2019.

Lempert, Robert J., Steven W. Popper, and Steven C. Bankes, *Shaping the Next One Hundred Years: New Methods for Quantitative, Long-Term Policy Analysis*, RAND Corporation, MR-1626-RPC, 2003. As of March 4, 2025: https://www.rand.org/pubs/monograph_reports/MR1626.html

Lentzos, Filippa, and Cédric Invernizzi, "Laboratories in the Cloud," Bulletin of the Atomic Scientists, July 2, 2019.

Li, Heng, and Richard Durbin, "Inference of Human Population History from Individual Whole-Genome Sequences," *Nature*, Vol. 475, July 28, 2011.

Los Alamos National Laboratory, "What Affects Effects?" April 2, 2024.

Lynch, Michael, John Conery, and Reinhold Burger, "Mutation Accumulation and the Extinction of Small Populations," *American Naturalist*, Vol. 146, No. 4, October 1995.

Lyons, Shelby L., Allison T. Karp, Timothy J. Bralower, Kliti Grice, Bettina Schaefer, Sean P. S. Gulick, Joanna V. Morgan, and Katherine H. Freeman, "Organic Matter from the Chicxulub Crater Exacerbated the K–Pg Impact Winter," *Proceedings of the National Academy of Sciences*, Vol. 117, No. 41, October 13, 2020.

MacAskill, William, *What We Owe the Future*, Basic Books, 2022.

Marcellino, William, Nathan Beauchamp-Mustafaga, Amanda Kerrigan, Lev Navarre Chao, and Jackson Smith, *The Rise of Generative AI and the Coming Era of Social Media Manipulation 3.0: Next-Generation Chinese Astroturfing and Coping with Ubiquitous AI,* RAND Corporation, PE-A2679-1, September 2023. As of March 18, 2025: https://www.rand.org/pubs/perspectives/PEA2679-1.html

Marchau, Vincent A. W. J., Warren E. Walker, Pieter J. T. M. Bloemen, and Steven W. Popper, eds., *Decision Making Under Deep Uncertainty: From Theory to Practice*, Springer Cham, 2019.

Marshall, I. D., and G. W. Douglas, “Studies in the Epidemiology of Infectious Myxomatosis of Rabbits: VIII. Further Observations on Changes in the Innate Resistance of Australian Wild Rabbits Exposed to Myxomatosis,” *Journal of Hygiene*, Vol. 59, No. 1, March 1961.

Matheny, Jason G., “Reducing the Risk of Human Extinction,” *Risk Analysis*, Vol. 27, No. 5, October 2007.

Matthews, Tom, Michael Byrne, Radley Horton, Conor Murphy, Roger Pielke Sr., Colin Raymond, Peter Thorne, and Robert L. Wilby, “Latent Heat Must Be Visible in Climate Communications,” *WIREs Climate Change*, Vol. 13, No. 4, July/August 2022.

Miller, B. R., M. Rigby, L. J. M. Kuijpers, P. B. Krummel, L. P. Steele, M. Leist, P. J. Fraser, A. McCulloch, C. Harth, P. Salameh, et al., “HFC-23 (CHF_3) Emission Trend Response to HCFC-22 (CHC_1F_2) Production and Recent HFC-23 Emission Abatement Measures,” *Atmospheric Chemistry and Physics*, Vol. 10, No. 16, August 25, 2010.

Millett, Piers, and Andrew Snyder-Beattie, “Existential Risk and Cost-Effective Biosecurity,” *Health Security*, Vol. 15, No. 4, July/August 2017.

Montreal Protocol on Substances That Deplete the Ozone Layer, September 16, 1987.

Montzka, Stephen A., Geoff S. Dutton, Pengfei Yu, Eric Ray, Robert W. Portmann, John S. Daniel, Lambert Kuijpers, Brad D. Hall, Debra Mondeel, Carolina Siso, et al., “An Unexpected and Persistent Increase in Global Emissions of Ozone-Depleting CFC-11,” *Nature*, Vol. 557, May 17, 2018.

Morgan, Joanna V., Timothy J. Bralower, Julia Brugger, and Kai Wünnemann, “The Chicxulub Impact and Its Environmental Consequences,” *Nature Reviews Earth and Environment*, Vol. 3, No. 5, May 2022.

Morgan, M. Granger, “Use (and Abuse) of Expert Elicitation in Support of Decision Making for Public Policy,” *Proceedings of the National Academy of Sciences*, Vol. 111, No. 20, May 20, 2014.

Mouton, Christopher A., Caleb Lucas, and Ella Guest, *The Operational Risks of AI in Large-Scale Biological Attacks: Results of a Red-Team Study*, RAND Corporation, RR-A2977-2, 2024. As of March 4, 2025: https://www.rand.org/pubs/research_reports/RRA2977-2.html

Ngo, Richard, “AGI Safety from First Principles: Control,” LessWrong, October 2, 2020.

Nguyen, Nhan, and Sarah Nadi, “An Empirical Evaluation of GitHub Copilot’s Code Suggestions,” *Proceedings of the 2022 Mining Software Repositories Conference: MSR 2022*, 2022.

NOAA Research, “Emissions of a Banned Ozone-Depleting Gas Are Back on the Decline,” February 10, 2021.

NOAA Research, “Two Additional Regions of Asia Were Sources of Banned Ozone-Destroying Chemicals,” March 9, 2022.

Office of the Deputy Assistant Secretary of Defense for Nuclear Matters, *Nuclear Matters Handbook 2020*, rev. ed., 2020.

Ogden, Pierce J., Eric D. Kelsic, Sam Sinai, and George M. Church, “Comprehensive AAV Capsid Fitness Landscape Reveals a Viral Gene and Enables Machine-Guided Design,” *Science*, Vol. 366, No. 6469, November 29, 2019.

Omohundro, Stephen M., “The Basic AI Drives,” *Artificial General Intelligence 2008: Proceedings of the First AGI Conference*, IOS Press, 2008.

Onyeaka, Helen, Christian K. Anumudu, Zainab T. Al-Sharify, Esther Egele-Godswill, and Paul Mbaegbu, “COVID-19 Pandemic: A Review of the Global Lockdown and Its Far-Reaching Effects,” *Science Progress*, Vol. 104, No. 2, April 2021.

Ord, Toby, *The Precipice: Existential Risk and the Future of Humanity*, Grand Central Publishing, 2020.

Organisation for Economic Co-operation and Development, *Recommendation of the Council on Artificial Intelligence*, OECD/LEGAL/0449, May 22, 2019.

Parker, Andy, and Peter J. Irvine, “The Risk of Termination Shock from Solar Geoengineering,” *Earth’s Future*, Vol. 6, No. 3, March 2018.

Pittock, A. B., T. P. Ackerman, P. J. Crutzen, M. C. MacCracken, C. S. Shapiro, and R. P. Turco, *Environmental Consequences of Nuclear War*: Vol. 1, *Physical and Atmospheric Effects*, John Wiley and Sons, 1986.

Poldrack, Russell A., Thomas Lu, and Gašper Beguš, "AI-Assisted Coding: Experiments with GPT-4," arXiv, arXiv:2304.13187, April 25, 2023.

Pope, F. D., P. Braesicke, R. G. Grainger, M. Kalberer, I. M. Watson, P. J. Davidson, and R. A. Cox, "Stratospheric Aerosol Particles and Solar-Radiation Management," *Nature Climate Change*, Vol. 2, No. 10, October 2012.

Popper, Steven W., "Reflections: DMDU and Public Policy for Uncertain Times," in Vincent A. W. J. Marchau, Warren E. Walker, Pieter J. T. M. Bloemen, and Steven W. Popper, eds., *Decision Making Under Deep Uncertainty: From Theory to Practice*, Springer Cham, 2019.

Reisner, Jon, Gennaro D'Angelo, Eunmo Koo, Wesley Even, Matthew Hecht, Elizabeth Hunke, Darin Comeau, Randall Bos, and James Cooley, "Climate Impact of a Regional Nuclear Weapons Exchange: An Improved Assessment Based on Detailed Source Calculations," *Journal of Geophysical Research: Atmospheres*, Vol. 123, No. 5, March 16, 2018.

Rekacewicz, Philippe, "Global CFC Production," webpage, GRID-Arendal, 2005. As of May 15, 2024: https://www.grida.no/resources/5506

Richards, C. E., H. L. Gauch, and J. M. Allwood, "International Risk of Food Insecurity and Mass Mortality in a Runaway Global Warming Scenario," *Futures*, Vol. 150, June 2023.

Richards, C. E., R. C. Lupton, and J. M. Allwood, "Re-Framing the Threat of Global Warming: An Empirical Causal Loop Diagram of Climate Change, Food Insecurity, and Societal Collapse," *Climatic Change*, Vol. 164, No. 49, 2021.

Robock, Alan, "Nuclear Winter," WIREs Climate Change, Vol. 1, No. 3, May/June 2010.

Robock, Alan, Luke Oman, and Georgiy L. Stenchikov, "Nuclear Winter Revisited with a Modern Climate Model and Current Nuclear Arsenals: Still Catastrophic Consequences," *Journal of Geophysical Research: Atmospheres*, Vol. 112, No. D13, July 16, 2007.

Rouse, Barry T., and Sharvan Sehrawat, "Immunity and Immunopathology to Viruses: What Decides the Outcome?" *Nature Reviews Immunology*, Vol. 10, No. 7, July 2010.

Sandbrink, Jonas B., "Artificial Intelligence and Biological Misuse: Differentiating Risks of Language Models and Biological Design Tools," arXiv, arXiv:2306.13952v8, last revised December 23, 2023.

Sanger, David E., "Putin's Move on Nuclear Treaty May Signal End to Formal Arms Control," *New York Times*, February 21, 2023.

Sanjuán, Rafael, Miguel R. Nebot, Nicola Chirico, Louis M. Mansky, and Robert Belshaw, "Viral Mutation Rates," *Journal of Virology*, Vol. 84, No. 19, October 2010.

Savelka, Jaromir, Arav Agarwal, Marshall An, Chris Bogart, and Majd Sakr, "Thrilled by Your Progress! Large Language Models (GPT-4) No Longer Struggle to Pass Assessments in Higher Education Programming Courses," *ICER '23: Proceedings of the 2023 ACM Conference on International Computing Education Research*, Vol. 1, September 2023.

Say, Daniel, Alistair J. Manning, Luke M. Western, Dickon Young, Adam Wisher, Matthew Rigby, Stefan Reimann, Martin K. Vollmer, Michela Maione, Jgor Arduini, et al., "Global Trends and European Emissions of Tetrafluoromethane (CF_4), Hexafluoroethane (C_2F_6), and Octafluoropropane (C_3F_8)," *Atmospheric Chemistry and Physics*, Vol. 21, No. 3, February 12, 2021.

Scouras, James, "Nuclear War as a Global Catastrophic Risk," *Journal of Benefit-Cost Analysis*, Vol. 10, No. 2, Summer 2019.

Scroxton, Alex, "Research Team Tricks AI Chatbots into Writing Usable Malicious Code," *Computer Weekly*, October 24, 2023.

Shapiro, Charles S., Ted F. Harvey, and Kendall R. Peterson, "Radioactive Fallout," in Fredric Solomon and Robert Q. Marston, eds., *The Medical Implications of Nuclear War*, National Academies Press, 1986.

Sherwood, Steven C., and Matthew Huber, "An Adaptability Limit to Climate Change Due to Heat Stress," *Proceedings of the National Academy of Sciences*, Vol. 107, No. 21, May 25, 2010.

Shevlane, Toby, Sebastian Farquhar, Ben Garfinkel, Mary Phuong, Jess Whittlestone, Jade Leung, Daniel Kokotajlo, Nahema Marchal, Markus Anderljung, Noam Kolt, et al., "Model Evaluation for Extreme Risks," arXiv, arXiv:2305.15324v2, September 22, 2023.

Shipman, Pat Lee, "The Bright Side of the Black Death," *American Scientist*, Vol. 102, No. 6, November–December 2014.

Shulman, Carl, *Omohundro's "Basic AI Drives" and Catastrophic Risks*, Singularity Institute, 2010.

Snyder, Don, Sarah A. Nowak, Mahyar A. Amouzegar, Julie Kim, and Richard Mesic, *Sustaining the U.S. Air Force Nuclear Mission*, RAND Corporation, TR-1240-AF, 2013. As of March 18, 2025: https://www.rand.org/pubs/technical_reports/TR1240.html

Soice, Emily H., Rafael Rocha, Kimberlee Cordova, Michael Specter, and Kevin M. Esvelt, "Can Large Language Models Democratize Access to Dual-Use Biotechnology?" arXiv, arXiv:2306.03809, June 6, 2023.

Statista, "Annual Production of Plastics Worldwide from 1950 to 2023," webpage, November 2024. As of May 15, 2024: https://www.statista.com/statistics/282732/global-production-of-plastics-since-1950/

Stauffer, Maxime, Konrad Seifert, Angela Aristizábal, Hamza Tariq Chaudhry, Kevin Kohler, Sumaya Nur Hussein, Claudette Salinas Levya, Arne Gebert, Jacob Arbeid, Mahaut Estier, et al., *Existential Risk and Rapid Technological Change: Advancing Risk-Informed Development*, United Nations Office for Disaster Risk Reduction, 2023.

Sunstein, Cass R., "Knightian Uncertainty," SSRN, December 11, 2023.

"Superforecasting Power-Seeking AI," *Good Judgment* blog, October 11, 2023.

Toon, Owen B., Charles G. Bardeen, Alan Robock, Lili Xia, Hans Kristensen, Matthew McKinzie, R. J. Peterson, Cheryl S. Harrison, Nicole S. Lovenduski, and Richard P. Turco, "Rapidly Expanding Nuclear Arsenals in Pakistan and India Portend Regional and Global Catastrophe," *Science Advances*, Vol. 5, No. 10, October 11, 2019.

Toon, O. B., R. P. Turco, A. Robock, C. Bardeen, L. Oman, and G. L. Stenchikov, "Atmospheric Effects and Societal Consequences of Regional Scale Nuclear Conflicts and Acts of Individual Nuclear Terrorism," *Atmospheric Chemistry and Physics*, Vol. 7, No. 8, April 19, 2007.

Traill, Lochran W., Barry W. Brook, Richard R. Frankham, and Corey J. A. Bradshaw, "Pragmatic Population Viability Targets in a Rapidly Changing World," *Biological Conservation*, Vol. 143, No. 1, January 2010.

Tsai, Wen-Tien, "Environmental and Health Risk Analysis of Nitrogen Trifluoride (NF_3), a Toxic and Potent Greenhouse Gas," *Journal of Hazardous Materials*, Vol. 159, Nos. 2–3, November 30, 2008.

Turchin, Alexey, and Brian Patrick Green, "Islands as Refuges for Surviving Global Catastrophes," *Foresight*, Vol. 21, No. 1, 2019.

Turco, R. P., O. B. Toon, T. P. Ackerman, J. B. Pollack, and Carl Sagan, "Nuclear Winter: Global Consequences of Multiple Nuclear Explosions," *Science*, Vol. 222, No. 4630, December 23, 1983.

UN—*See* United Nations.

United Nations, *The World's Cities in 2016*, September 2016.

U.S. Antarctic Program, "About the Continent," webpage, undated. As of April 12, 2024: https://www.usap.gov/aboutthecontinent/

U.S. Department of Defense, *Military and Security Developments Involving the People's Republic of China*, 2023.

U.S. Environmental Protection Agency, "Radioactive Fallout from Nuclear Weapons Testing," webpage, last updated May 29, 2024. As of March 3, 2025: https://www.epa.gov/radtown/radioactive-fallout-nuclear-weapons-testing

U.S. Environmental Protection Agency, "Understanding Global Warming Potentials," webpage, last updated January 16, 2025. As of March 3, 2025: https://www.epa.gov/ghgemissions/understanding-global-warming-potentials

Vecellio, Daniel J., Qinqin Kong, W. Larry Kenney, and Matthew Huber, "Greatly Enhanced Risk to Humans as a Consequence of Empirically Determined Lower Moist Heat Stress Tolerance," *Proceedings of the National Academy of Sciences*, Vol. 120, No. 42, October 17, 2023.

Vecellio, Daniel J., S. Tony Wolf, Rachel M. Cottle, and W. Larry Kenney, “Evaluating the 35°C Wet-Bulb Temperature Adaptability Threshold for Young, Healthy Subjects (PSU HEAT Project),” *Journal of Applied Physiology*, Vol. 132, No. 2, February 2022.

Wagman, Benjamin M., Katherine A. Lundquist, Qi Tang, Lee G. Glascoe, and David C. Bader, “Examining the Climate Effects of a Regional Nuclear Weapons Exchange Using a Multiscale Atmospheric Modeling Approach,” *Journal of Geophysical Research: Atmospheres*, Vol. 125, No. 24, December 27, 2020.

Walker, Warren E., Vincent A. W. J. Marchau, and Jan H. Kwakkel, “Dynamic Adaptive Planning (DAP),” in Vincent A. W. J. Marchau, Warren E. Walker, Pieter J. T. M. Bloemen, and Steven W. Popper, eds., *Decision Making Under Deep Uncertainty: From Theory to Practice*, Springer Cham, 2019.

Watson, Joseph L., David Juergens, Nathaniel R. Bennett, Brian L. Trippe, Jason Yim, Helen E. Eisenach, Woody Ahern, Andrew J. Borst, Robert J. Ragotte, Lukas F. Milles, et al., “De Novo Design of Protein Structure and Function with RFdiffusion,” *Nature*, Vol. 620, No. 7976, August 31, 2023.

Weiss, Ray. F, Jens Mühle, Peter K. Salameh, and Christina M. Harth, “Nitrogen Trifluoride in the Global Atmosphere,” *Atmospheric Science*, Vol. 35, No. 20, October 2008.

Willis, Henry H., Anu Narayanan, Benjamin Boudreaux, Bianca Espinosa, Edward Geist, Daniel M. Gerstein, Dahlia Anne Goldfeld, Nidhi Kalra, Tom LaTourrette, Emily Lathrop, Alvin Moon, Jan Osburg, Benjamin Lee Preston, Kristin Van Abel, Emmi Yonekura, Robert J. Lempert, Sunny D. Bhatt, Chandra Garber, and Emily Lawson, *Global Catastrophic Risk Assessment,* Homeland Security Operational Analysis Center operated by the RAND Corporation, RR-A2981-1, 2024. As of March 5, 2025:
https://www.rand.org/pubs/research_reports/RRA2981-1.html

Witze, Alexandra, “How a Small Nuclear War Would Transform the Entire Planet,” *Nature*, Vol. 579, No. 7800, March 16, 2020.

Wyatt, Kelly B., Paula F. Campos, M. Thomas P. Gilbert, Sergios-Orestis Kolokotronis, Wayne H. Hynes, Rob DeSalle, Peter Daszak, Ross D. E. MacPhee, and Alex D. Greenwood, “Historical Mammal Extinction on Christmas Island (Indian Ocean) Correlates with Introduced Infectious Disease,” *PLOS One*, Vol. 3, No. 11, 2008.

Xia, Lili, Alan Robock, Kim Scherrer, Cheryl S. Harrison, Benjamin Leon Bodirsky, Isabelle Weindl, Jonas Jägermeyr, Charles G. Bardeen, Owen B. Toon, and Ryan Heneghan, “Global Food Insecurity and Famine from Reduced Crop, Marine Fishery and Livestock Production Due to Climate Disruption from Nuclear War Soot Injection,” *Nature Food*, Vol. 3, No. 8, August 2022.

Xu, Chi, Timothy A. Kohler, Timothy M. Lenton, Jens-Christian Svenning, and Marten Scheffer, “Future of the Human Climate Niche,” *Proceedings of the National Academy of Sciences*, Vol. 117, No. 21, May 26, 2020.

Xu, Yangyang, and Veerabhadran Ramanathan, “Well Below 2°C: Mitigation Strategies for Avoiding Dangerous to Catastrophic Climate Changes,” *Proceedings of the National Academy of Sciences*, Vol. 114, No. 39, September 26, 2017.

Xu, Yangyang, Nathanael Philip Ribar, Gunnar W. Schade, Andrew John Lockley, Yi Ge Zhang, Jeffrey Sachnik, Pengfei Yu, Jianxin Hu, and Guus J. M. Velders, “Possible Mitigation of Global Cooling Due to Supervolcanic Eruption via Intentional Release of Fluorinated Gases,” ESS Open Archive, May 13, 2023.

Yudkowsky, Eliezer, *Intelligence Explosion Microeconomics*, Machine Intelligence Research Institute, September 13, 2013.

Yudkowsky, Eliezer, “AGI Ruin: A List of Lethalities,” AI Alignment Forum, June 5, 2022.

About the Authors

Michael J. D. Vermeer is a senior physical scientist at RAND. He researches science and technology policy in homeland security, the armed forces, the intelligence community, and criminal justice while focusing on technology governance and national security implications of emerging technologies. He holds a Ph.D. in inorganic chemistry.

Emily Lathrop is an associate engineer at RAND. Her research at RAND covers topics including AI, human-machine teaming, robotics, and unmanned systems. She holds a Ph.D. in mechanical engineering.

Alvin Moon is an associate mathematician at RAND. His research at RAND covers topics including AI, cryptography, supply chains, and workforce development, with a focus on analysis and modeling. He has a Ph.D. in mathematics.