CHAPTER 1

Overview of Catastrophic AI Risks

1.1 INTRODUCTION

In this chapter, we will give a brief and informal description of many major societal-scale risks from AI, focusing on AI risks that could lead to highly severe or even catastrophic societal outcomes. This provides some background and motivation before we discuss specific challenges with more depth and rigor in the following chapters.

The world as we know it today is not normal. We take for granted that we can talk instantaneously with people thousands of miles away, fly to the other side of the world in less than a day, and access vast mountains of accumulated knowledge on devices we carry around in our pockets. These realities seemed far-fetched decades ago, and would have been inconceivable to people living centuries ago. The ways we live, work, travel, and communicate have only been possible for a tiny fraction of human history.

Yet, when we look at the bigger picture, a broader pattern emerges: accelerating development. Hundreds of thousands of years elapsed between the time Homo sapiens appeared on Earth and the agricultural revolution. Then, thousands of years passed before the industrial revolution. Now, just centuries later, the artificial intelligence (AI) revolution is beginning. The march of history is not constant—it is rapidly accelerating.

We can capture this trend quantitatively in Figure 1.1, which shows how estimated gross world product has changed over time [3, 4]. The hyperbolic growth it depicts might be explained by the fact that, as technology advances, the rate of technological advancement also tends to increase. Empowered with new technologies, people can innovate faster than they could before. Thus, the gap in time between each landmark development narrows.

It is the rapid pace of development, as much as the sophistication of our technology, that makes the present day an unprecedented time in human history. We have

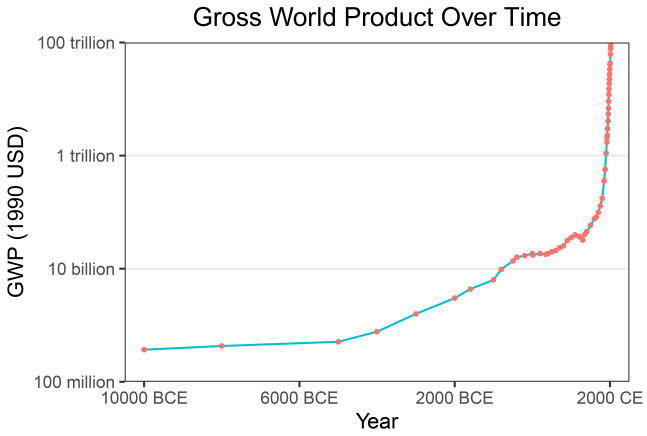

FIGURE 1.1. World production has grown rapidly over the course of human history. AI could further this trend, catapulting humanity into a new period of unprecedented change.

reached a point where technological advancements can transform the world beyond recognition within a human lifetime. For example, people who have lived through the creation of the internet can remember a time when our now digitally-connected world would have seemed like science fiction.

From a historical perspective, it appears possible that the same amount of development could now be condensed in an even shorter timeframe. We might not be certain that this will occur, but neither can we rule it out. We therefore wonder: what new technology might usher in the next big acceleration? In light of recent advances, AI seems an increasingly plausible candidate. Perhaps, as AI continues to become more powerful, it could lead to a qualitative shift in the world, more profound than any we have experienced so far. It could be the most impactful period in history, though it could also be the last.

Although technological advancement has often improved people's lives, we ought to remember that, as our technology grows in power, so too does its destructive potential. Consider the invention of nuclear weapons. Last century, for the first time in our species' history, humanity possessed the ability to destroy itself, and the world suddenly became much more fragile.

Our newfound vulnerability revealed itself in unnerving clarity during the Cold War. On a Saturday in October 1962, the Cuban Missile Crisis was cascading out of control. US warships enforcing the blockade of Cuba detected a Soviet submarine and attempted to force it to the surface by dropping low-explosive depth charges. The submarine was out of radio contact, and its crew had no idea whether World War III had already begun. A broken ventilator raised the temperature up to 140°F in some parts of the submarine, causing crew members to fall unconscious as depth charges exploded nearby.

The submarine carried a nuclear-armed torpedo, which required consent from both the captain and political officer to launch. Both provided it. On any other submarine in Cuban waters that day, that torpedo would have launched—and a nuclear third

world war may have followed. Fortunately, a man named Vasili Arkhipov was also on the submarine. Arkhipov was the commander of the entire flotilla and by sheer luck happened to be on that particular submarine. He talked the captain down from his rage, convincing him to await further orders from Moscow. He averted a nuclear war and saved millions or billions of lives—and possibly civilization itself.

Carl Sagan once observed, "If we continue to accumulate only power and not wisdom, we will surely destroy ourselves" [5]. Sagan was correct: The power of nuclear weapons was not one we were ready for. Overall, it has been luck rather than wisdom that has saved humanity from nuclear annihilation, with multiple recorded instances of a single individual preventing a full-scale nuclear war.

AI is now poised to become a powerful technology with destructive potential similar to nuclear weapons. We do not want to repeat the Cuban Missile Crisis. We do not want to slide toward a moment of peril where our survival hinges on luck rather than the ability to use this technology wisely. Instead, we need to work proactively to mitigate the risks it poses. This necessitates a better understanding of what could go wrong and what to do about it.

Luckily, AI systems are not yet advanced enough to contribute to every risk we discuss. But that is cold comfort in a time when AI development is advancing at an unprecedented and unpredictable rate. We consider risks arising from both present-day AIs and AIs that are likely to exist in the near future. It is possible that if we wait for more advanced systems to be developed before taking action, it may be too late.

In this chapter, we will explore various ways in which powerful AIs could bring about catastrophic events with devastating consequences for vast numbers of people. We will also discuss how AIs could present existential risks—catastrophes from which humanity would be unable to recover. The most obvious such risk is extinction, but there are other outcomes, such as creating a permanent dystopian society, which would also constitute an existential catastrophe. As further discussed in this book's Introduction, we do not intend to cover all risks or harms that AI may pose in an exhaustive manner, and many of these fall outside the scope of this chapter. We outline many possible scenarios, some of which are more likely than others and some of which are mutually incompatible with each other. This approach is motivated by the principles of risk management. We prioritize asking "what could go wrong?" rather than reactively waiting for catastrophes to occur. This proactive mindset enables us to anticipate and mitigate catastrophic risks before it's too late.

To help orient the discussion, we decompose catastrophic risks from AIs into four risk sources that warrant intervention:

- **Malicious use**: Malicious actors using AIs to cause large-scale devastation.
- **AI race**: Competitive pressures that could drive us to deploy AIs in unsafe ways, despite this being in no one's best interest.
- **Organizational risks**: Accidents arising from the complexity of AIs and the organizations developing them.

- **Rogue AIs**: The problem of controlling a technology more intelligent than we are.

These four sections—Malicious Use, AI Race, Organizational Risks, and Rogue AIs—describe causes of AI risks that are *intentional*, *environmental/structural*, *accidental*, and *internal*, respectively [6]. The risks that are briefly outlined in this chapter are discussed in greater depth in the rest of this book.

In this chapter, we will describe how concrete, small-scale examples of each risk might escalate into catastrophic outcomes. We also include hypothetical stories to help readers conceptualize the various processes and dynamics discussed in each section. We hope this survey will serve as a practical introduction for readers interested in learning about and mitigating catastrophic AI risks.

1.2 MALICIOUS USE

On the morning of March 20, 1995, five men entered the Tokyo subway system. After boarding separate subway lines, they continued for several stops before dropping the bags they were carrying and exiting. An odorless, colorless liquid inside the bags began to vaporize. Within minutes, commuters began choking and vomiting. The trains continued on toward the heart of Tokyo, with sickened passengers leaving the cars at each station. The fumes were spread at each stop, either by emanating from the tainted cars or through contact with people's clothing and shoes. By the end of the day, 13 people lay dead and 5,800 seriously injured. The group responsible for the attack was the religious cult Aum Shinrikyo [7]. Its motive for murdering innocent people? To bring about the end of the world.

Powerful new technologies offer tremendous potential benefits, but they also carry the risk of empowering malicious actors to cause widespread harm. There will always be those with the worst of intentions, and AIs could provide them with a formidable tool to achieve their objectives. Moreover, as AI technology advances, severe malicious use could potentially destabilize society, increasing the likelihood of other risks.

In this section, we will explore the various ways in which the malicious use of advanced AIs could pose catastrophic risks. These include engineering biochemical weapons, unleashing rogue AIs, using persuasive AIs to spread propaganda and erode consensus reality, and leveraging censorship and mass surveillance to irreversibly concentrate power. We will conclude by discussing possible strategies for mitigating the risks associated with the malicious use of AIs.

Unilateral actors considerably increase the risks of malicious use. In instances where numerous actors have access to a powerful technology or dangerous information that could be used for harmful purposes, it only takes one individual to cause significant devastation. Malicious actors themselves are the clearest example of this, but recklessness can be equally dangerous. For example, a single research team might be excited to open source an AI system with biological research capabilities,

which would speed up research and potentially save lives, but this could also increase the risk of malicious use if the AI system could be repurposed to develop bioweapons. In situations like this, the outcome may be determined by the least risk-averse research group. If only one research group thinks the benefits outweigh the risks, it could act unilaterally, deciding the outcome even if most others don't agree. And if they are wrong and someone does decide to develop a bioweapon, it would be too late to reverse course.

By default, advanced AIs may increase the destructive capacity of both the most powerful and the general population. Thus, the growing potential for AIs to empower malicious actors is one of the most severe threats humanity will face in the coming decades. The examples we give in this section are only those we can foresee. It is possible that AIs could aid in the creation of dangerous new technology we cannot presently imagine, which would further increase risks from malicious use.

1.2.1 Bioterrorism

The rapid advancement of AI technology increases the risk of bioterrorism. AIs with knowledge of bioengineering could facilitate the creation of novel bioweapons and lower barriers to obtaining such agents. Engineered pandemics from AI-assisted bioweapons pose a unique challenge, as attackers have an advantage over defenders and could constitute an existential threat to humanity. We will now examine these risks and how AIs might exacerbate challenges in managing bioterrorism and engineered pandemics.

Bioengineered pandemics present a new threat. Biological agents, including viruses and bacteria, have caused some of the most devastating catastrophes in history. It's believed the Black Death killed more humans than any other event in history, an astounding and awful 200 million, the equivalent to four billion deaths today. While contemporary advancements in science and medicine have made great strides in mitigating risks associated with natural pandemics, engineered pandemics could be designed to be more lethal or easily transmissible than natural pandemics, presenting a new threat that could equal or even surpass the devastation wrought by history's most deadly plagues [8].

Humanity has a long and dark history of weaponizing pathogens, with records dating back to 1320 BCE describing a war in Asia Minor where infected sheep were driven across the border to spread Tularemia [9]. During the twentieth century, 15 countries are known to have developed bioweapons programs, including the US, USSR, UK, and France. Like chemical weapons, bioweapons have become a taboo among the international community. While some state actors continue to operate bioweapons programs [10], a more significant risk may come from non-state actors like Aum Shinrikyo, ISIS, or simply disturbed individuals. Due to advancements in AI and biotechnology, the tools and knowledge necessary to engineer pathogens with capabilities far beyond Cold War-era bioweapons programs will rapidly democratize.

Biotechnology is progressing rapidly and becoming more accessible. A few decades ago, the ability to synthesize new viruses was limited to a handful of the top scientists working in advanced laboratories. Today it is estimated that there are 30,000 people with the talent, training, and access to technology to create new pathogens [8]. This figure could rapidly expand. Gene synthesis, which allows the creation of custom biological agents, has dropped precipitously in price, with its cost halving approximately every 15 months [11]. Furthermore, with the advent of benchtop DNA synthesis machines, access will become much easier and could avoid existing gene synthesis screening efforts, which complicates controlling the spread of such technology [12]. The chances of a bioengineered pandemic killing millions, perhaps billions, is proportional to the number of people with the skills and access to the technology to synthesize them. With AI assistants, orders of magnitude more people could have the required skills, thereby increasing the risks by orders of magnitude.

AIs could be used to expedite the discovery of new, more deadly chemical and biological weapons. In 2022, researchers took an AI system designed to create new drugs by generating non-toxic, therapeutic molecules and tweaked it to reward, rather than penalize, toxicity [13]. After this simple change, within six hours, it generated 40,000 candidate chemical warfare agents entirely on its own. It designed not just known deadly chemicals including VX, but also novel molecules that may be deadlier than any chemical warfare agents discovered so far. In the field of biology, AIs have already surpassed human abilities in protein structure prediction [14] and made contributions to synthesizing those proteins [15]. Similar methods could be used to create bioweapons and develop pathogens that are deadlier, more transmissible, and more difficult to treat than anything seen before.

AIs compound the threat of bioengineered pandemics. AIs will increase the number of people who could commit acts of bioterrorism. General-purpose AIs like ChatGPT are capable of synthesizing expert knowledge about the deadliest known pathogens, such as influenza and smallpox, and providing step-by-step instructions about how a person could create them while evading safety protocols [16]. Future versions of AIs could be even more helpful to potential bioterrorists when AIs are able to synthesize information into techniques, processes, and knowledge that is not explicitly available anywhere on the internet. Public health authorities may respond to these threats with safety measures, but in bioterrorism, the attacker has the advantage. The exponential nature of biological threats means that a single attack could spread to the entire world before an effective defense could be mounted. Only 100 days after being detected and sequenced, the omicron variant of COVID-19 had infected a quarter of the United States and half of Europe [8]. Quarantines and lockdowns instituted to suppress the COVID-19 pandemic caused a global recession and still could not prevent the disease from killing millions worldwide.

In summary, advanced AIs could constitute a weapon of mass destruction in the hands of terrorists, by making it easier for them to design, synthesize, and spread deadly new pathogens. By reducing the required technical expertise and increasing the lethality and transmissibility of pathogens, AIs could enable malicious actors to cause global catastrophe by unleashing pandemics.

1.2.2 Unleashing AI Agents

Many technologies are *tools* that humans use to pursue our goals, such as hammers, toasters, and toothbrushes. But AIs are increasingly built as *agents* which autonomously take actions in the world in order to pursue open-ended goals. AI agents can be given goals such as winning games, making profits on the stock market, or driving a car to a destination. AI agents therefore pose a unique risk: people could build AIs that pursue dangerous goals.

Malicious actors could intentionally create rogue AIs. One month after the release of GPT-4, an open-source project bypassed the AI's safety filters and turned it into an autonomous AI agent instructed to "destroy humanity," "establish global dominance," and "attain immortality." Dubbed ChaosGPT, the AI compiled research on nuclear weapons and sent tweets trying to influence others. Fortunately, ChaosGPT was merely a warning given that it lacked the ability to successfully formulate long-term plans, hack computers, and survive and spread. Yet given the rapid pace of AI development, ChaosGPT did offer a glimpse into the risks that more advanced rogue AIs could pose in the near future.

Many groups may want to unleash AIs or have AIs displace humanity. Simply unleashing rogue AIs, like a more sophisticated version of ChaosGPT, could accomplish mass destruction, even if those AIs aren't explicitly told to harm humanity. There are a variety of beliefs that may drive individuals and groups to do so. One ideology that could pose a unique threat in this regard is "accelerationism." This ideology seeks to accelerate AI development as rapidly as possible and opposes restrictions on the development or proliferation of AIs. This sentiment is common among many leading AI researchers and technology leaders, some of whom are intentionally racing to build AIs more intelligent than humans. According to Google co-founder Larry Page, AIs are humanity's rightful heirs and the next step of cosmic evolution. He has also expressed the sentiment that humans maintaining control over AIs is "speciesist" [17]. Jürgen Schmidhuber, an eminent AI scientist, argued that "In the long run, humans will not remain the crown of creation... But that's okay because there is still beauty, grandeur, and greatness in realizing that you are a tiny part of a much grander scheme which is leading the universe from lower complexity towards higher complexity" [18]. Richard Sutton, another leading AI scientist, in discussing smarter-than human AI asked "why shouldn't those who are the smartest become powerful?" and thinks the development of superintelligence will be an achievement "beyond humanity, beyond life, beyond good and bad" [19]. He argues that "succession to AI is inevitable," and while "they could displace us from existence," "we should not resist succession" [20].

There are several sizable groups who may want to unleash AIs to intentionally cause harm. For example, sociopaths and psychopaths make up around 3 percent of the population [21]. In the future, people who have their livelihoods destroyed by AI automation may grow resentful, and some may want to retaliate. There are plenty of cases in which seemingly mentally stable individuals with no history of insanity or violence suddenly go on a shooting spree or plant a bomb with the intent to

harm as many innocent people as possible. We can also expect well-intentioned people to make the situation even more challenging. As AIs advance, they could make ideal companions—knowing how to provide comfort, offering advice when needed, and never demanding anything in return. Inevitably, people will develop emotional bonds with chatbots, and some will demand that they be granted rights or become autonomous.

In summary, releasing powerful AIs and allowing them to take actions independently of humans could lead to a catastrophe. There are many reasons that people might pursue this, whether because of a desire to cause harm, an ideological belief in technological acceleration, or a conviction that AIs should have the same rights and freedoms as humans.

1.2.3 Persuasive AIs

The deliberate propagation of disinformation is already a serious issue, reducing our shared understanding of reality and polarizing opinions. AIs could be used to severely exacerbate this problem by generating personalized disinformation on a larger scale than before. Additionally, as AIs become better at predicting and nudging our behavior, they will become more capable at manipulating us. We will now discuss how AIs could be leveraged by malicious actors to create a fractured and dysfunctional society.

AIs could pollute the information ecosystem with motivated lies. Sometimes ideas spread not because they are true, but because they serve the interests of a particular group. "Yellow journalism" was coined as a pejorative reference to newspapers that advocated war between Spain and the United States in the late 19th century, because they believed that sensational war stories would boost their sales [22]. When public information sources are flooded with falsehoods, people will sometimes fall prey to lies, or else come to distrust mainstream narratives, both of which undermine societal integrity.

Unfortunately, AIs could escalate these existing problems dramatically. First, AIs could be used to generate unique, personalized disinformation at a large scale. While there are already many social media bots [23], some of which exist to spread disinformation, historically they have been run by humans or primitive text generators. The latest AI systems do not need humans to generate personalized messages, never get tired, and could potentially interact with millions of users at once [24].

AIs can exploit users' trust. Already, hundreds of thousands of people pay for chatbots marketed as lovers and friends [25], and one man's suicide has been partially attributed to interactions with a chatbot [26]. As AIs appear increasingly human-like, people will increasingly form relationships with them and grow to trust them. AIs that gather personal information through relationship-building or by accessing extensive personal data, such as a user's email account or personal files, could leverage that information to enhance persuasion. Powerful actors that control those systems could

exploit user trust by delivering personalized disinformation directly through people's "friends."

AIs could centralize control of trusted information. Separate from democratizing disinformation, AIs could centralize the creation and dissemination of trusted information. Only a few actors have the technical skills and resources to develop cutting-edge AI systems, and they could use these AIs to spread their preferred narratives. Alternatively, if AIs are broadly accessible this could lead to widespread disinformation, with people retreating to trusting only a small handful of authoritative sources [27]. In both scenarios, there would be fewer sources of trusted information and a small portion of society would control popular narratives.

AI censorship could further centralize control of information. This could begin with good intentions, such as using AIs to enhance fact-checking and help people avoid falling prey to false narratives. This would not necessarily solve the problem, as disinformation persists today despite the presence of fact-checkers.

Even worse, purported "fact-checking AIs" might be designed by authoritarian governments and others to suppress the spread of true information. Such AIs could be designed to correct most common misconceptions but provide incorrect information about some sensitive topics, such as human rights violations committed by certain countries. But even if fact-checking AIs work as intended, the public might eventually become entirely dependent on them to adjudicate the truth, reducing people's autonomy and making them vulnerable to failures or hacks of those systems.

In a world with widespread persuasive AI systems, people's beliefs might be almost entirely determined by which AI systems they interact with most. Never knowing whom to trust, people could retreat even further into ideological enclaves, fearing that any information from outside those enclaves might be a sophisticated lie. This would erode consensus reality, people's ability to cooperate with others, participate in civil society, and address collective action problems. This would also reduce our ability to have a conversation as a species about how to mitigate existential risks from AIs.

In summary, AIs could create highly effective, personalized disinformation on an unprecedented scale, and could be particularly persuasive to people they have built personal relationships with. In the hands of many people, this could create a deluge of disinformation that debilitates human society, but, kept in the hands of a few, it could allow governments to control narratives for their own ends.

1.2.4 Concentration of Power

We have discussed several ways in which individuals and groups might use AIs to cause widespread harm, through bioterrorism; releasing powerful, uncontrolled AIs; and disinformation. To mitigate these risks, governments might pursue intense surveillance and seek to keep AIs in the hands of a trusted minority. This reaction, however, could easily become an overcorrection, paving the way for an entrenched totalitarian

regime that would be locked in by the power and capacity of AIs. This scenario represents a form of "top-down" misuse, as opposed to "bottom-up" misuse by citizens, and could in extreme cases culminate in an entrenched dystopian civilization.

AIs could lead to extreme, and perhaps irreversible concentration of power. The persuasive abilities of AIs combined with their potential for surveillance and the advancement of autonomous weapons could allow small groups of actors to "lock-in" their control over society, perhaps permanently. To operate effectively, AIs require a broad set of infrastructure components, which are not equally distributed, such as data centers, computing power, and big data. Those in control of powerful systems may use them to suppress dissent, spread propaganda and disinformation, and otherwise advance their goals, which may be contrary to public wellbeing.

AIs may entrench a totalitarian regime. In the hands of the state, AIs may result in the erosion of civil liberties and democratic values in general. AIs could allow totalitarian governments to efficiently collect, process, and act on an unprecedented volume of information, permitting an ever smaller group of people to surveil and exert complete control over the population without the need to enlist millions of citizens to serve as willing government functionaries. Overall, as power and control shift away from the public and toward elites and leaders, democratic governments are highly vulnerable to totalitarian backsliding. Additionally, AIs could make totalitarian regimes much longer-lasting; a major way in which such regimes have been toppled previously is at moments of vulnerability like the death of a dictator, but AIs, which would be hard to "kill," could provide much more continuity to leadership, providing few opportunities for reform.

AIs can entrench corporate power at the expense of the public good. Corporations have long lobbied to weaken laws and policies that restrict their actions and power, all in the service of profit. Corporations in control of powerful AI systems may use them to manipulate customers into spending more on their products even to the detriment of their own wellbeing. The concentration of power and influence that could be afforded by AIs could enable corporations to exert unprecedented control over the political system and entirely drown out the voices of citizens. This could occur even if creators of these systems know their systems are self-serving or harmful to others, as they would have incentives to reinforce their power and avoid distributing control.

In addition to power, locking in certain values may curtail humanity's moral progress. It's dangerous to allow any set of values to become permanently entrenched in society. For example, AI systems have learned racist and sexist views [28], and once those views are learned, it can be difficult to fully remove them. In addition to problems we know exist in our society, there may be some we still do not. Just as we abhor some moral views widely held in the past, people in the future may want to move past moral views that we hold today, even those we currently see no problem with. For example, moral defects in AI systems would be even worse if AI systems had been trained in the 1960s, and many people at the time would have seen

no problem with that. We may even be unknowingly perpetuating moral catastrophes today [29]. Therefore, when advanced AIs emerge and transform the world, there is a risk of their objectives locking in or perpetuating defects in today's values. If AIs are not designed to continuously learn and update their understanding of societal values, they may perpetuate or reinforce existing defects in their decision-making processes long into the future.

In summary, although keeping powerful AIs in the hands of a few might reduce the risks of terrorism, it could further exacerbate power inequality if misused by governments and corporations. This could lead to totalitarian rule and intense manipulation of the public by corporations, and could lock in current values, preventing any further moral progress.

Story: Bioterrorism

The following is an illustrative hypothetical story to help readers envision some of these risks. This story is nonetheless somewhat vague to reduce the risk of inspiring malicious actions based on it.

A biotechnology startup is making waves in the industry with its AI-powered bioengineering model. The company has made bold claims that this new technology will revolutionize medicine through its ability to create cures for both known and unknown diseases. The company did, however, stir up some controversy when it decided to release the program to approved researchers in the scientific community. Only weeks after its decision to make the model open-source on a limited basis, the full model was leaked on the internet for all to see. Its critics pointed out that the model could be repurposed to design lethal pathogens and claimed that the leak provided bad actors with a powerful tool to cause widespread destruction, opening it up to abuse without safeguards in place.

Unknown to the public, an extremist group has been working for years to engineer a new virus designed to kill large numbers of people. Yet given their lack of expertise, these efforts have so far been unsuccessful. When the new AI system is leaked, the group immediately recognizes it as a potential tool to design the virus and circumvent legal and monitoring obstacles to obtain the necessary raw materials. The AI system successfully designs exactly the kind of virus the extremist group was hoping for. It also provides step-by-step instructions on how to synthesize large quantities of the virus and circumvent any obstacles to spreading it. With the synthesized virus in hand, the extremist group devises a plan to release the virus in several carefully chosen locations in order to maximize its spread.

The virus has a long incubation period and spreads silently and quickly throughout the population for months. By the time it is detected, it has already infected millions and has an alarmingly high mortality rate. Given its lethality, most who are infected will ultimately die. The virus may or may not be contained eventually, but not before it kills millions of people.

1.3 AI RACE

The immense potential of AIs has created competitive pressures among global players contending for power and influence. This "AI race" is driven by nations and corporations who feel they must rapidly build and deploy AIs to secure their positions and survive. By failing to properly prioritize global risks, this dynamic makes it more likely that AI development will produce dangerous outcomes. Analogous to the nuclear arms race during the Cold War, participation in an AI race may serve individual short-term interests, but it ultimately results in worse collective outcomes for humanity. Importantly, these risks stem not only from the intrinsic nature of AI technology, but from the competitive pressures that encourage insidious choices in AI development.

In this section, we first explore the military AI arms race and the corporate AI race, where nation-states and corporations are forced to rapidly develop and adopt AI systems to remain competitive. Moving beyond these specific races, we reconceptualize competitive pressures as part of a broader evolutionary process in which AIs could become increasingly pervasive, powerful, and entrenched in society. Finally, we highlight potential strategies and policy suggestions to mitigate the risks created by an AI race and ensure the safe development of AIs.

1.3.1 Military AI Arms Race

The development of AIs for military applications is swiftly paving the way for a new era in military technology, with potential consequences rivaling those of gunpowder and nuclear arms in what has been described as the "third revolution in warfare." The weaponization of AI presents numerous challenges, such as the potential for more destructive wars, the possibility of accidental usage or loss of control, and the prospect of malicious actors co-opting these technologies for their own purposes. As AIs gain influence over traditional military weaponry and increasingly take on command and control functions, humanity faces a paradigm shift in warfare. In this context, we will discuss the latent risks and implications of this AI arms race on global security, the potential for intensified conflicts, and the dire outcomes that could come as a result, including the possibility of conflicts escalating to a scale that poses an existential threat.

Lethal Autonomous Weapons (LAWs)

LAWs are weapons that can identify, target, and kill without human intervention [30]. They offer potential improvements in decision-making speed and precision. Warfare, however, is a high-stakes, safety-critical domain for AIs with significant moral and practical concerns. Though their existence is not necessarily a catastrophe in itself, LAWs may serve as an on-ramp to catastrophes stemming from malicious use, accidents, loss of control, or an increased likelihood of war.

LAWs may become vastly superior to humans. Driven by rapid developments in AIs, weapons systems that can identify, target, and decide to kill human beings on their own—without an officer directing an attack or a soldier pulling the trigger—are starting to transform the future of conflict. In 2020, an advanced AI agent outperformed experienced F-16 pilots in a series of virtual dogfights, including decisively defeating a human pilot 5-0, showcasing "aggressive and precise maneuvers the human pilot couldn't outmatch" [31]. Just as in the past, superior weapons would allow for more destruction in a shorter period of time, increasing the severity of war.

Militaries are taking steps toward delegating life-or-death decisions to AIs. Fully autonomous drones were likely first used on the battlefield in Libya in March 2020, when retreating forces were "hunted down and remotely engaged" by a drone operating without human oversight [32]. In May 2021, the Israel Defense Forces used the world's first AI-guided weaponized drone swarm during combat operations, which marks a significant milestone in the integration of AI and drone technology in warfare [33]. Although walking, shooting robots have yet to replace soldiers on the battlefield, technologies are converging in ways that may make this possible in the near future.

LAWs increase the likelihood of war. Sending troops into battle is a grave decision that leaders do not make lightly. But autonomous weapons would allow an aggressive nation to launch attacks without endangering the lives of its own soldiers and thus face less domestic scrutiny. While remote-controlled weapons share this advantage, their scalability is limited by the requirement for human operators and vulnerability to jamming countermeasures, limitations that LAWs could overcome [34]. Public opinion for continuing wars tends to wane as conflicts drag on and casualties increase [35]. LAWs would change this equation. National leaders would no longer face the prospect of body bags returning home, thus removing a primary barrier to engaging in warfare, which could ultimately increase the likelihood of conflicts.

Cyberwarfare

As well as being used to enable deadlier weapons, AIs could lower the barrier to entry for cyberattacks, making them more numerous and destructive. They could cause serious harm not only in the digital environment but also in physical systems, potentially taking out critical infrastructure that societies depend on. While AIs could also be used to improve cyberdefense, it is unclear whether they will be most effective as an offensive or defensive technology [36]. If they enhance attacks more than they support defense, then cyberattacks could become more common, creating significant geopolitical turbulence and paving another route to large-scale conflict.

AIs have the potential to increase the accessibility, success rate, scale, speed, stealth, and potency of cyberattacks. Cyberattacks are already a reality, but AIs could be used to increase their frequency and destructiveness in multiple ways. Machine learning tools could be used to find more critical vulnerabilities in target systems and improve the success rate of attacks. They could also be used to

increase the scale of attacks by running millions of systems in parallel, and increase the speed by finding novel routes to infiltrating a system. Cyberattacks could also become more potent if used to hijack AI weapons.

Cyberattacks can destroy critical infrastructure. By hacking computer systems that control physical processes, cyberattacks could cause extensive infrastructure damage. For example, they could cause system components to overheat or valves to lock, leading to a buildup of pressure culminating in an explosion. Through interferences like this, cyberattacks have the potential to destroy critical infrastructure, such as electric grids and water supply systems. This was demonstrated in 2015, when a cyberwarfare unit of the Russian military hacked into the Ukrainian power grid, leaving over 200,000 people without power access for several hours. AI-enhanced attacks could be even more devastating and potentially deadly for the billions of people who rely on critical infrastructure for survival.

Difficulties in attributing AI-driven cyberattacks could increase the risk of war. A cyberattack resulting in physical damage to critical infrastructure would require a high degree of skill and effort to execute, perhaps only within the capability of nation-states. Such attacks are rare as they constitute an act of war, and thus elicit a full military response. Yet AIs could enable attackers to hide their identity, for example if they are used to evade detection systems or more effectively cover the tracks of the attacker [37]. If cyberattacks become more stealthy, this would reduce the threat of retaliation from an attacked party, potentially making attacks more likely. If stealthy attacks do happen, they might incite actors to mistakenly retaliate against unrelated third parties they suspect to be responsible. This could increase the scope of the conflict dramatically.

Automated Warfare

AIs speed up the pace of war, which makes AIs more necessary. AIs can quickly process a large amount of data, analyze complex situations, and provide helpful insights to commanders. With ubiquitous sensors and advanced technology on the battlefield, there is tremendous incoming information. AIs help make sense of this information, spotting important patterns and relationships that humans might miss. As these trends continue, it will become increasingly difficult for humans to make well-informed decisions as quickly as necessary to keep pace with AIs. This would further pressure militaries to hand over decisive control to AIs. The continuous integration of AIs into all aspects of warfare will cause the pace of combat to become faster and faster. Eventually, we may arrive at a point where humans are no longer capable of assessing the ever-changing battlefield situation and must cede decision-making power to advanced AIs.

Automatic retaliation can escalate accidents into war. There is already willingness to let computer systems retaliate automatically. In 2014, a leak revealed to the public that the NSA was developing a system called MonsterMind, which would autonomously detect and block cyberattacks on US infrastructure [38]. It was

suggested that in the future, MonsterMind could automatically initiate a retaliatory cyberattack with no human involvement. If multiple combatants have policies of automatic retaliation, an accident or false alarm could quickly escalate to full-scale war before humans intervene. This would be especially dangerous if the superior information processing capabilities of modern AI systems makes it more appealing for actors to automate decisions regarding nuclear launches.

History shows the danger of automated retaliation. On September 26, 1983, Stanislav Petrov, a lieutenant colonel of the Soviet Air Defense Forces, was on duty at the Serpukhov-15 bunker near Moscow, monitoring the Soviet Union's early warning system for incoming ballistic missiles. The system indicated that the US had launched multiple nuclear missiles toward the Soviet Union. The protocol at the time dictated that such an event should be considered a legitimate attack, and the Soviet Union would respond with a nuclear counterstrike. If Petrov had passed on the warning to his superiors, this would have been the likely outcome. Instead, however, he judged it to be a false alarm and ignored it. It was soon confirmed that the warning had been caused by a rare technical malfunction. If an AI had been in control, the false alarm could have triggered a nuclear war.

AI-controlled weapons systems could lead to a flash war. Autonomous systems are not infallible. We have already witnessed how quickly an error in an automated system can escalate in the economy. Most notably, in the 2010 Flash Crash, a feedback loop between automated trading algorithms amplified ordinary market fluctuations into a financial catastrophe in which a trillion dollars of stock value vanished in minutes [39]. If multiple nations were to use AIs to automate their defense systems, an error could be catastrophic, triggering a spiral of attacks and counter-attacks that would happen too quickly for humans to step in—a flash war. The market quickly recovered from the 2010 Flash Crash, but the harm caused by a flash war could be catastrophic.

Automated warfare could reduce accountability for military leaders. Military leaders may at times gain an advantage on the battlefield if they are willing to ignore the laws of war. For example, soldiers may be able to mount stronger attacks if they do not take steps to minimize civilian casualties. An important deterrent to this behavior is the risk that military leaders could eventually be held accountable or even prosecuted for war crimes. Automated warfare could reduce this deterrence effect by making it easier for military leaders to escape accountability by blaming violations on failures in their automated systems.

AIs could make war more uncertain, increasing the risk of conflict. Although states that are already wealthier and more powerful often have more resources to invest in new military technologies, they are not necessarily always the most successful at adopting them. Other factors also play an important role, such as how agile and adaptive a military can be in incorporating new technologies [40]. Major new weapons innovations can therefore offer an opportunity for existing superpowers to bolster their dominance, but also for less powerful states to quickly increase their

power by getting ahead in an emerging and important sphere. This can create significant uncertainty around if and how the balance of power is shifting, potentially leading states to incorrectly believe they could gain something from going to war. Even aside from considerations regarding the balance of power, rapidly evolving automated warfare would be unprecedented, making it difficult for actors to evaluate their chances of victory in any particular conflict. This would increase the risk of miscalculation, making war more likely.

Actors May Risk Extinction Over Individual Defeat

Competitive pressures make actors more willing to accept the risk of extinction. During the Cold War, neither side desired the dangerous situation they found themselves in. There were widespread fears that nuclear weapons could be powerful enough to wipe out a large fraction of humanity, potentially even causing extinction—a catastrophic result for both sides. Yet the intense rivalry and geopolitical tensions between the two superpowers fueled a dangerous cycle of arms buildup. Each side perceived the other's nuclear arsenal as a threat to its very survival, leading to a desire for parity and deterrence. The competitive pressures pushed both countries to continually develop and deploy more advanced and destructive nuclear weapons systems, driven by the fear of being at a strategic disadvantage. During the Cuban Missile Crisis, this led to the brink of nuclear war. Even though the story of Arkhipov preventing the launch of a nuclear torpedo wasn't declassified until decades after the incident, President John F. Kennedy reportedly estimated that he thought the odds of nuclear war beginning during that time were "somewhere between one out of three and even." This chilling admission highlights how the competitive pressures between militaries have the potential to cause global catastrophes.

Individually rational decisions can be collectively catastrophic. Nations locked in competition might make decisions that advance their own interests by putting the rest of the world at stake. Scenarios of this kind are collective action problems, where decisions may be rational on an individual level yet disastrous for the larger group [41]. For example, corporations and individuals may weigh their own profits and convenience over the negative impacts of the emissions they create, even if those emissions collectively result in climate change. The same principle can be extended to military strategy and defense systems. Military leaders might estimate, for instance, that increasing the autonomy of weapon systems would mean a 10 percent chance of losing control over weaponized superhuman AIs. Alternatively, they might estimate that using AIs to automate bioweapons research could lead to a 10 percent chance of leaking a deadly pathogen. Both of these scenarios could lead to catastrophe or even extinction. The leaders may, however, also calculate that refraining from these developments will mean a 99 percent chance of losing a war against an opponent. Since conflicts are often viewed as existential struggles by those fighting them, rational actors may accept an otherwise unthinkable 10 percent chance of human extinction over a 99 percent chance of losing a war. Regardless of the particular nature of the risks posed by advanced AIs, these dynamics could push us to the brink of global catastrophe.

Technological superiority does not guarantee national security. It is tempting to think that the best way of guarding against enemy attacks is to improve one's own military prowess. However, in the midst of competitive pressures, all parties will tend to advance their weaponry, such that no one gains much of an advantage, but all are left at greater risk. As Richard Danzig, former Secretary of the Navy, has observed, "The introduction of complex, opaque, novel, and interactive technologies will produce accidents, emergent effects, and sabotage. On a number of occasions and in a number of ways, the American national security establishment will lose control of what it creates... deterrence is a strategy for reducing attacks, not accidents" [42].

Cooperation is paramount to reducing risk. As discussed above, an AI arms race can lead us down a hazardous path, despite this being in no country's best interest. It is important to remember that we are all on the same side when it comes to existential risks, and working together to prevent them is a collective necessity. A destructive AI arms race benefits nobody, so all actors would be rational to take steps to cooperate with one another to prevent the riskiest applications of militarized AIs. As Dwight D. Eisenhower reminded us, "The only way to win World War III is to prevent it."

We have considered how competitive pressures could lead to the increasing automation of conflict, even if decision-makers are aware of the existential threat that this path entails. We have also discussed cooperation as being the key to counteracting and overcoming this collective action problem. We will now illustrate a hypothetical path to disaster that could result from an AI arms race.

Story: Automated Warfare

As AI systems become increasingly sophisticated, militaries start involving them in decision-making processes. Officials give them military intelligence about opponents' arms and strategies, for example, and ask them to calculate the most promising plan of action. It soon becomes apparent that AIs are reliably reaching better decisions than humans, so it seems sensible to give them more influence. At the same time, international tensions are rising, increasing the threat of war.

A new military technology has recently been developed that could make international attacks swifter and stealthier, giving targets less time to respond. Since military officials feel their response processes take too long, they fear that they could be vulnerable to a surprise attack capable of inflicting decisive damage before they would have any chance to retaliate. Since AIs can process information and make decisions much more quickly than humans, military leaders reluctantly hand them increasing amounts of retaliatory control, reasoning that failing to do so would leave them open to attack from adversaries.

While for years military leaders had stressed the importance of keeping a "human in the loop" for major decisions, human control is nonetheless gradually

phased out in the interests of national security. Military leaders understand that their decisions lead to the possibility of inadvertent escalation caused by system malfunctions, and would prefer a world where all countries automated less; but they do not trust that their adversaries will refrain from automation. Over time, more and more of the chain of command is automated on all sides.

One day, a single system malfunctions, detecting an enemy attack when there is none. The system is empowered to launch an instant "retaliatory" attack, and it does so in the blink of an eye. The attack causes automated retaliation from the other side, and so on. Before long, the situation is spiraling out of control, with waves of automated attack and retaliation. Although humans have made mistakes leading to escalation in the past, this escalation between mostly-automated militaries happens far more quickly than any before. The humans who are responding to the situation find it difficult to diagnose the source of the problem, as the AI systems are not transparent. By the time they even realize how the conflict started, it is already over, with devastating consequences for both sides.

1.3.2 Corporate AI Race

Competitive pressures exist in the economy, as well as in military settings. Although competition between companies can be beneficial, creating more useful products for consumers, there are also pitfalls. First, the benefits of economic activity may be unevenly distributed, incentivizing those who benefit most from it to disregard the harms to others. Second, under intense market competition, businesses tend to focus much more on short-term gains than on long-term outcomes. With this mindset, companies often pursue something that can make a lot of profit in the short term, even if it poses a societal risk in the long term. We will now discuss how corporate competitive pressures could play out with AIs and the potential negative impacts.

Economic Competition Undercuts Safety

Competitive pressure is fueling a corporate AI race. To obtain a competitive advantage, companies often race to offer the first products to a market rather than the safest. These dynamics are already playing a role in the rapid development of AI technology. At the launch of Microsoft's AI-powered search engine in February 2023, the company's CEO Satya Nadella said, "A race starts today... we're going to move fast." Only weeks later, the company's chatbot was shown to have threatened to harm users [43]. In an internal email, Sam Schillace, a technology executive at Microsoft, highlighted the urgency in which companies view AI development. He wrote that it would be an "absolutely fatal error in this moment to worry about things that can be fixed later" [44].

Competitive pressures have contributed to major commercial and industrial disasters. Throughout the 1960s, Ford Motor Company faced competition

from international car manufacturers as the share of imports in American car purchases steadily rose [45]. Ford developed an ambitious plan to design and manufacture a new car model in only 25 months [46]. The Ford Pinto was delivered to customers ahead of schedule, but with a serious safety problem: the gas tank was located near the rear bumper, and could explode during rear collisions. Numerous fatalities and injuries were caused by the resulting fires when crashes inevitably happened [47]. Ford was sued and a jury found them liable for these deaths and injuries [48]. The verdict, of course, came too late for those who had already lost their lives. As Ford's president at the time was fond of saying, "Safety doesn't sell" [49].

Boeing, aiming to compete with its rival Airbus, sought to deliver an updated, more fuel-efficient model to the market as quickly as possible. The head-to-head rivalry and time pressure led to the introduction of the Maneuvering Characteristics Augmentation System, which was designed to enhance the aircraft's stability. However, inadequate testing and pilot training ultimately resulted in the two fatal crashes only months apart, with 346 people killed [50]. We can imagine a future in which similar pressures lead companies to cut corners and release unsafe AI systems.

A third example is the Bhopal gas tragedy, which is widely considered to be the worst industrial disaster ever to have happened. In December 1984, a vast quantity of toxic gas leaked from a Union Carbide Corporation subsidiary plant manufacturing pesticides in Bhopal, India. Exposure to the gas killed thousands of people and injured up to half a million more. Investigations found that, in the run-up to the disaster, safety standards had fallen significantly, with the company cutting costs by neglecting equipment maintenance and staff training as profitability fell. This is often considered a consequence of competitive pressures [51].

Competition incentivizes businesses to deploy potentially unsafe AI systems. In an environment where businesses are rushing to develop and release products, those that follow rigorous safety procedures will be slower and risk being outcompeted. Ethically-minded AI developers, who want to proceed more cautiously and slow down, would give more unscrupulous developers an advantage. In trying to survive commercially, even the companies that want to take more care are likely to be swept along by competitive pressures. There may be attempts to implement safety measures, but with more of an emphasis on capabilities than on safety, these may be insufficient. This could lead us to develop highly powerful AIs before we properly understand how to ensure they are safe.

Automated Economy

Corporations will face pressure to replace humans with AIs. As AIs become more capable, they will be able to perform an increasing variety of tasks more quickly, cheaply, and effectively than human workers. Companies will therefore stand to gain a competitive advantage from replacing their employees with AIs. Companies that choose not to adopt AIs would likely be out-competed, just as a clothing company using manual looms would be unable to keep up with those using industrial ones.

AIs could lead to mass unemployment. Economists have long considered the possibility that machines will replace human labor. Nobel Prize winner Wassily Leontief said in 1952 that, as technology advances, "Labor will become less and less important... more and more workers will be replaced by machines" [52]. Previous technologies have augmented the productivity of human labor. AIs, however, could differ profoundly from previous innovations. Advanced AIs capable of automating human labor should be regarded not merely as tools, but as agents. Human-level AI agents would, by definition, be able to do everything a human could do. These AI agents would also have important advantages over human labor. They could work 24 hours a day, be copied many times and run in parallel, and process information much more quickly than a human would. While we do not know when this will occur, it is unwise to discount the possibility that it could be soon. If human labor is replaced by AIs, mass unemployment could dramatically increase inequality, making individuals dependent on the owners of AI systems.

Automated AI R&D. AI agents would have the potential to automate the research and development (R&D) of AI itself. AI is increasingly automating parts of the research process [53], and this could lead to AI capabilities growing at increasing rates, to the point where humans are no longer the driving force behind AI development. If this trend continues unchecked, it could escalate risks associated with AIs progressing faster than our capacity to manage and regulate them. Imagine that we created an AI that writes and thinks at the speed of today's AIs, but that it could also perform world-class AI research. We could then copy that AI and create 10,000 world-class AI researchers that operate at a pace 100× times faster than humans. By automating AI research and development, we might achieve progress equivalent to many decades in just a few months.

Conceding power to AIs could lead to human enfeeblement. Even if we ensure that the many unemployed humans are provided for, we may find ourselves completely reliant on AIs. This would likely emerge not from a violent coup by AIs, but from a gradual slide into dependence. As society's challenges become ever more complex and fast-paced, and as AIs become ever more intelligent and quick-thinking, we may forfeit more and more functions to them out of convenience. In such a state, the only feasible solution to the complexities and challenges compounded by AIs may be to rely even more heavily on AIs. This gradual process could eventually lead to the delegation of nearly all intellectual, and eventually physical, labor to AIs. In such a world, people might have few incentives to gain knowledge and cultivate skills, potentially leading to a state of enfeeblement [54]. Having lost our know-how and our understanding of how civilization works, we would become completely dependent on AIs, a scenario not unlike the one depicted in the film WALL-E. In such a state, humanity is not flourishing and is no longer in effective control.

As we have seen, there are classic game-theoretic dilemmas where individuals and groups face incentives that are incompatible with what would make everyone better off. We see this with a military AI arms race, where the world is made less safe by creating extremely powerful AI weapons, and we see this in a corporate AI race,

where an AI's power and development is prioritized over its safety. To address these dilemmas that give rise to global risks, we will need new coordination mechanisms and institutions. It is our view that failing to coordinate and stop AI races would be the most likely cause of an existential catastrophe.

1.3.3 Evolutionary Pressures

As discussed above, there are strong pressures to replace humans with AIs, cede more control to them, and reduce human oversight in various settings, despite the potential harms. We can re-frame this as a general trend resulting from evolutionary dynamics, where an unfortunate truth is that AIs will simply be more fit than humans. Extrapolating this pattern of automation, it is likely that we will build an ecosystem of competing AIs over which it may be difficult to maintain control in the long run. We will now discuss how natural selection influences the development of AI systems and why evolution favors selfish behaviors. We will also look at how competition might arise and play out between AIs and humans, and how this could create catastrophic risks. This section draws heavily from "*Natural Selection Favors AIs over Humans*" [55, 56].

Fitter technologies are selected, for good and bad. While most people think of evolution by natural selection as a biological process, its principles shape much more. According to the evolutionary biologist Richard Lewontin [57], evolution by natural selection will take hold in any environment where three conditions are present: 1) there are differences between individuals; 2) characteristics are passed onto future generations and; 3) the different variants propagate at different rates. These conditions apply to various technologies.

Consider the content-recommendation algorithms used by streaming services and social media platforms. When a particularly addictive content format or algorithm hooks users, it results in higher screen time and engagement. This more effective content format or algorithm is consequently "selected" and further fine-tuned, while formats and algorithms that fail to capture attention are discontinued. These competitive pressures foster a "survival of the most addictive" dynamic. Platforms that refuse to use addictive formats and algorithms become less influential or are simply outcompeted by platforms that do, leading competitors to undermine wellbeing and cause massive harm to society [58].

The conditions for natural selection apply to AIs. There will be many different AI developers who make many different AI systems with varying features and capabilities, and competition between them will determine which characteristics become more common. Second, the most successful AIs today are already being used as a basis for their developers' next generation of models, as well as being imitated by rival companies. Third, factors determining which AIs propagate the most may include their ability to act autonomously, automate labor, or reduce the chance of their own deactivation.

Natural selection often favors selfish characteristics. Natural selection influences which AIs propagate most widely. From biological systems, we see that natural selection often gives rise to selfish behaviors that promote one's own genetic information: chimps attack other communities [59], lions engage in infanticide [60], viruses evolve new surface proteins to deceive and bypass defense barriers [61], humans engage in nepotism, some ants enslave others [62], and so on. In the natural world, selfishness often emerges as a dominant strategy; those that prioritize themselves and those similar to them are usually more likely to survive, so these traits become more prevalent. Amoral competition can select for traits that we think are immoral.

Examples of selfish behaviors. For concreteness, we now describe many selfish traits—traits that expand AIs' influence at the expense of humans. AIs that automate a task and leave many humans jobless have engaged in selfish behavior; these AIs may not even be aware of what a human is but still behave selfishly towards them—selfish behaviors do not require malicious intent. Likewise, AI managers may engage in selfish and "ruthless" behavior by laying off thousands of workers; such AIs may not even believe they did anything wrong—they were just being "efficient." AIs may eventually become enmeshed in vital infrastructure such as power grids or the internet. Many people may then be unwilling to accept the cost of being able to effortlessly deactivate them, as that would pose a reliability hazard. AIs that help create a new useful system—a company, or infrastructure—that becomes increasingly complicated and eventually requires AIs to operate them also have engaged in selfish behavior. AIs that help people develop AIs that are more intelligent—but happen to be less interpretable to humans—have engaged in selfish behavior, as this reduces human control over AIs' internals. AIs that are more charming, attractive, hilarious, imitate sentience (uttering phrases like "ouch!" or pleading "please don't turn me off!"), or emulate deceased family members are more likely to have humans grow emotional connections with them. These AIs are more likely to cause outrage at suggestions to destroy them, and they are more likely preserved, protected, or granted rights by some individuals. If some AIs are given rights, they may operate, adapt, and evolve outside of human control. Overall, AIs could become embedded in human society and expand their influence over us in ways that we can't reverse.

Selfish behaviors may erode safety measures that some of us implement. AIs that gain influence and provide economic value will predominate, while AIs that adhere to the most constraints will be less competitive. For example, AIs following the constraint "never break the law" have fewer options than AIs following the constraint "don't get caught breaking the law." AIs of the latter type may be willing to break the law if they're unlikely to be caught or if the fines are not severe enough, allowing them to outcompete more restricted AIs. Many businesses follow laws, but in situations where stealing trade secrets or deceiving regulators is highly lucrative and difficult to detect, a business that is willing to engage in such selfish behavior can have an advantage over its more principled competitors.

An AI system might be prized for its ability to achieve ambitious goals autonomously. It might, however, be achieving its goals efficiently without abiding by ethical re-

strictions, while deceiving humans about its methods. Even if we try to put safety measures in place, a deceptive AI would be very difficult to counteract if it is cleverer than us. AIs that can bypass our safety measures without detection may be the most successful at accomplishing the tasks we give them, and therefore become widespread. These processes could culminate in a world where many aspects of major companies and infrastructure are controlled by powerful AIs with selfish traits, including deceiving humans, harming humans in service of their goals, and preventing themselves from being deactivated.

Humans only have nominal influence over AI selection. One might think we could avoid the development of selfish behaviors by ensuring we do not select AIs that exhibit them. However, the companies developing AIs are not selecting the safest path but instead succumbing to evolutionary pressures. One example is OpenAI, which was founded as a nonprofit in 2015 to "benefit humanity as a whole, unconstrained by a need to generate financial return" [63]. However, when faced with the need to raise capital to keep up with better-funded rivals, in 2019 OpenAI transitioned from a nonprofit to "capped-profit" structure [64]. Later, many of the safety-focused OpenAI employees left and formed a competitor, Anthropic, that was to focus more heavily on AI safety than OpenAI had. Although Anthropic originally focused on safety research, they eventually became convinced of the "necessity of commercialization" and now contribute to competitive pressures [65]. While many of the employees at those companies genuinely care about safety, these values do not stand a chance against evolutionary pressures, which compel companies to move ever more hastily and seek ever more influence, lest the company perish. Moreover, AI developers are already selecting AIs with increasingly selfish traits. They are selecting AIs to automate and displace humans, make humans highly dependent on AIs, and make humans more and more obsolete. By their own admission, future versions of these AIs may lead to extinction [66]. This is why an AI race is insidious: AI development is not being aligned with human values but rather with natural selection.

People often choose the products that are most useful and convenient to them immediately, rather than thinking about potential long-term consequences, even to themselves. An AI race puts pressures on companies to select the AIs that are most competitive, not the least selfish. Even if it's feasible to select for unselfish AIs, if it comes at a clear cost to competitiveness, some competitors will select the selfish AIs. Furthermore, as we have mentioned, if AIs develop strategic awareness, they may counteract our attempts to select against them. Moreover, as AIs increasingly automate various processes, AIs will affect the competitiveness of other AIs, not just humans. AIs will interact and compete with each other, and some will be put in charge of the development of other AIs at some point. Giving AIs influence over which other AIs should be propagated and how they should be modified would represent another step toward humans becoming dependent on AIs and AI evolution becoming increasingly independent from humans. As this continues, the complex process governing AI evolution will become further unmoored from human interests.

AIs can be more fit than humans. Our unmatched intelligence has granted us power over the natural world. It has enabled us to land on the moon, harness nuclear energy, and reshape landscapes at our will. It has also given us power over other species. Although a single unarmed human competing against a tiger or gorilla has no chance of winning, the collective fate of these animals is entirely in our hands. Our cognitive abilities have proven so advantageous that, if we chose to, we could cause them to go extinct in a matter of weeks. Intelligence was a key factor that led to our dominance, but we are currently standing on the precipice of creating entities far more intelligent than ourselves.

Given the exponential increase in microprocessor speeds, AIs have the potential to process information and "think" at a pace that far surpasses human neurons, but it could be even more dramatic than the speed difference between humans and sloths—possibly more like the speed difference between humans and plants. They can assimilate vast quantities of data from numerous sources simultaneously, with near-perfect retention and understanding. They do not need to sleep and they do not get bored. Due to the scalability of computational resources, an AI could interact and cooperate with an unlimited number of other AIs, potentially creating a collective intelligence that would far outstrip human collaborations. AIs could also deliberately update and improve themselves. Without the same biological restrictions as humans, they could adapt and therefore evolve unspeakably quickly compared with us. Computers are becoming faster. Humans aren't [67].

To further illustrate the point, imagine that there was a new species of humans. They do not die of old age, they get 30% faster at thinking and acting each year, and they can instantly create adult offspring for the modest sum of a few thousand dollars. It seems clear, then, this new species would eventually have more influence over the future. In sum, AIs could become like an invasive species, with the potential to outcompete humans. Our only advantage over AIs is that we get to make the first moves, but given the frenzied AI race, we are rapidly giving up even this advantage.

AIs would have little reason to cooperate with or be altruistic toward humans. Cooperation and altruism evolved because they increase fitness. There are numerous reasons why humans cooperate with other humans, like direct reciprocity. Also known as "quid pro quo," direct reciprocity can be summed up by the idiom "you scratch my back, I'll scratch yours." While humans would initially select AIs that were cooperative, the natural selection process would eventually go beyond our control, once AIs were in charge of many or most processes, and interacting predominantly with one another. At that point, there would be little we could offer AIs, given that they will be able to "think" at least hundreds of times faster than us. Involving us in any cooperation or decision-making processes would simply slow them down, giving them no more reason to cooperate with us than we do with gorillas. It might be difficult to imagine a scenario like this or to believe we would ever let it happen. Yet it may not require any conscious decision, instead arising as we allow ourselves to gradually drift into this state without realizing that human-AI co-evolution may not turn out well for humans.

AIs becoming more powerful than humans could leave us highly vulnerable. As the most dominant species, humans have deliberately harmed many other species, and helped drive species such as woolly mammoths and Neanderthals to extinction. In many cases, the harm was not even deliberate, but instead a result of us merely prioritizing our goals over their wellbeing. To harm humans, AIs wouldn't need to be any more genocidal than someone removing an ant colony on their front lawn. If AIs are able to control the environment more effectively than we can, they could treat us with the same disregard.

Conceptual summary. Evolution could cause the most influential AI agents to act selfishly because:

1. **Evolution by natural selection gives rise to selfish behavior.** While evolution can result in altruistic behavior in rare situations, the context of AI development does not promote altruistic behavior.
2. **Natural selection may be a dominant force in AI development.** The intensity of evolutionary pressure will be high if AIs adapt rapidly or if competitive pressures are intense. Competition and selfish behaviors may dampen the effects of human safety measures, leaving the surviving AI designs to be selected naturally.

If so, AI agents would have many selfish tendencies. The winner of the AI race would not be a nation-state, not a corporation, but AIs themselves. The upshot is that the AI ecosystem would eventually stop evolving on human terms, and we would become a displaced, second-class species.

Story: Autonomous Economy

As AIs become more capable, people realize that we could work more efficiently by delegating some simple tasks to them, like drafting emails. Over time, people notice that the AIs are doing these tasks more quickly and effectively than any human could, so it is convenient to give them more jobs with less and less supervision.

Competitive pressures accelerate the expansion of AI use, as companies can gain an advantage over rivals by automating whole processes or departments with AIs, which perform better than humans and cost less to employ. Other companies, faced with the prospect of being out-competed, feel compelled to follow suit just to keep up. At this point, natural selection is already at work among AIs; humans choose to make more of the best-performing models and unwittingly propagate selfish traits such as deception and self-preservation if these confer a fitness advantage. For example, AIs that are charming and foster personal relationships with humans become widely copied and harder to remove.

As AIs are put in charge of more and more decisions, they are increasingly interacting with one another. Since they can evaluate information much more quickly than humans, activity in most spheres accelerates. This creates a feed-

back loop: since business and economic developments are too fast-moving for humans to follow, it makes sense to cede yet more control to AIs instead, pushing humans further out of important processes. Ultimately, this leads to a fully autonomous economy, governed by an increasingly uncontrolled ecosystem of AIs.

At this point, humans have few incentives to gain any skills or knowledge, because almost everything would be taken care of by much more capable AIs. As a result, we eventually lose the capacity to look after and govern ourselves. Additionally, AIs become convenient companions, offering social interaction without requiring the reciprocity or compromise necessary in human relationships. Humans interact less and less with one another over time, losing vital social skills and the ability to cooperate. People become so dependent on AIs that it would be intractable to reverse this process. What's more, as some AIs become more intelligent, some people are convinced these AIs should be given rights, meaning turning off some AIs is no longer a viable option.

Competitive pressures between the many interacting AIs continue to select for selfish behaviors, though we might be oblivious to this happening, as we have already acquiesced much of our oversight. If these clever, powerful, self-preserving AIs were then to start acting in harmful ways, it would be all but impossible to deactivate them or regain control.

AIs have supplanted humans as the most dominant species and their continued evolution is far beyond our influence. Their selfish traits eventually lead them to pursue their goals without regard for human wellbeing, with catastrophic consequences.

1.4 ORGANIZATIONAL RISKS

In January 1986, tens of millions of people tuned in to watch the launch of the Challenger Space Shuttle. Approximately 73 seconds after liftoff, the shuttle exploded, resulting in the deaths of everyone on board. Though tragic enough on its own, one of its crew members was a school teacher named Sharon Christa McAuliffe. McAuliffe was selected from over 10,000 applicants for the NASA Teacher in Space Project and was scheduled to become the first teacher to fly in space. As a result, millions of those watching were schoolchildren. NASA had the best scientists and engineers in the world, and if there was ever a mission NASA didn't want to go wrong, it was this one [68].

The Challenger disaster, alongside other catastrophes, serves as a chilling reminder that even with the best expertise and intentions, accidents can still occur. As we progress in developing advanced AI systems, it is crucial to remember that these systems are not immune to catastrophic accidents. An essential factor in preventing accidents and maintaining low levels of risk lies in the organizations responsible for

these technologies. In this section, we discuss how organizational safety plays a critical role in the safety of AI systems. First, we discuss how even without competitive pressures or malicious actors, accidents can happen—in fact, they are inevitable. We then discuss how improving organizational factors can reduce the likelihood of AI catastrophes.

Catastrophes occur even when competitive pressures are low. Even in the absence of competitive pressures or malicious actors, factors like human error or unforeseen circumstances can still bring about catastrophe. The Challenger disaster illustrates that organizational negligence can lead to loss of life, even when there is no urgent need to compete or outperform rivals. By January 1986, the space race between the US and USSR had largely diminished, yet the tragic event still happened due to errors in judgment and insufficient safety precautions.

Similarly, the Chernobyl nuclear disaster in April 1986 highlights how catastrophic accidents can occur in the absence of external pressures. As a state-run project without the pressures of international competition, the disaster happened when a safety test involving the reactor's cooling system was mishandled by an inadequately prepared night shift crew. This led to an unstable reactor core, causing explosions and the release of radioactive particles that contaminated large swathes of Europe [69]. Seven years earlier, America came close to experiencing its own Chernobyl when, in March 1979, a partial meltdown occurred at the Three Mile Island nuclear power plant. Though less catastrophic than Chernobyl, both events highlight how even with extensive safety measures in place and few outside influences, catastrophic accidents can still occur.

Another example of a costly lesson on organizational safety came just one month after the accident at Three Mile Island. In April 1979, spores of *Bacillus anthracis*—or simply "anthrax," as it is commonly known—were accidentally released from a Soviet military research facility in the city of Sverdlovsk. This led to an outbreak of anthrax that resulted in at least 66 confirmed deaths [70]. Investigations into the incident revealed that the cause of the release was a procedural failure and poor maintenance of the facility's biosecurity systems, despite being operated by the state and not subjected to significant competitive pressures.

The unsettling reality is that AI is far less understood and AI industry standards are far less stringent than nuclear technology and rocketry. Nuclear reactors are based on solid, well-established and well-understood theoretical principles. The engineering behind them is informed by that theory, and components are stress-tested to the extreme. Nonetheless, nuclear accidents still happen. In contrast, AI lacks a comprehensive theoretical understanding, and its inner workings remain a mystery even to those who create it. This presents an added challenge of controlling and ensuring the safety of a technology that we do not yet fully comprehend.

AI accidents could be catastrophic. Accidents in AI development could have devastating consequences. For example, imagine an organization unintentionally introduces a critical bug in an AI system designed to accomplish a specific task, such as

helping a company improve its services. This bug could drastically alter the AI's behavior, leading to unintended and harmful outcomes. One historical example of such a case occurred when researchers at OpenAI were attempting to train an AI system to generate helpful, uplifting responses. During a code cleanup, the researchers mistakenly flipped the sign of the reward used to train the AI [71]. As a result, instead of generating helpful content, the AI began producing hate-filled and sexually explicit text overnight without being halted. Accidents could also involve the unintentional release of a dangerous, weaponized, or lethal AI system. Since AIs can be easily duplicated with a simple copy-paste, a leak or hack could quickly spread the AI system beyond the original developers' control. Once the AI system becomes publicly available, it would be nearly impossible to put the genie back in the bottle.

Gain-of-function research could potentially lead to accidents by pushing the boundaries of an AI system's destructive capabilities. In these situations, researchers might intentionally train an AI system to be harmful or dangerous in order to understand its limitations and assess possible risks. While this can lead to useful insights into the risks posed by a given AI system, future gain-of-function research on advanced AIs might uncover capabilities significantly worse than anticipated, creating a serious threat that is challenging to mitigate or control. As with viral gain-of-function research, pursuing AI gain-of-function research may only be prudent when conducted with strict safety procedures, oversight, and a commitment to responsible information sharing. These examples illustrate how AI accidents could be catastrophic and emphasize the crucial role that organizations developing these systems play in preventing such accidents.

1.4.1 Accidents Are Hard to Avoid

When dealing with complex systems, the focus needs to be placed on ensuring accidents don't cascade into catastrophes. In his book "*Normal Accidents: Living with High-Risk Technologies*," sociologist Charles Perrow argues that accidents are inevitable and even "normal" in complex systems, as they are not merely caused by human errors but also by the complexity of the systems themselves [72]. In particular, such accidents are likely to occur when the intricate interactions between components cannot be completely planned or foreseen. For example, in the Three Mile Island accident, a contributing factor to the lack of situational awareness by the reactor's operators was the presence of a yellow maintenance tag, which covered valve position lights in the emergency feedwater lines [73]. This prevented operators from noticing that a critical valve was closed, demonstrating the unintended consequences that can arise from seemingly minor interactions within complex systems.

Unlike nuclear reactors, which are relatively well-understood despite their complexity, complete technical knowledge of most complex systems is often nonexistent. This is especially true of deep learning systems, for which the inner workings are exceedingly difficult to understand, and where the reason why certain design choices work can be hard to understand even in hindsight. Furthermore, unlike components in other industries, such as gas tanks, which are highly reliable, deep learning systems are

neither perfectly accurate nor highly reliable. Thus, the focus for organizations dealing with complex systems, especially deep learning systems, should not be solely on eliminating accidents, but rather on ensuring that accidents do not cascade into catastrophes.

Accidents are hard to avoid because of sudden, unpredictable developments. Scientists, inventors, and experts often significantly underestimate the time it takes for a groundbreaking technological advancement to become a reality. The Wright brothers famously claimed that powered flight was fifty years away, just two years before they achieved it. Lord Rutherford, a prominent physicist and the father of nuclear physics, dismissed the idea of extracting energy from nuclear fission as "moonshine," only for Leo Szilard to invent the nuclear chain reaction less than 24 hours later. Similarly, Enrico Fermi expressed 90 percent confidence in 1939 that it was impossible to use uranium to sustain a fission chain reaction—yet, just four years later he was personally overseeing the first reactor [74].

AI development could catch us off guard too. In fact, it often does. The defeat of Lee Sedol by AlphaGo in 2016 came as a surprise to many experts, as it was widely believed that achieving such a feat would still require many more years of development. More recently, large language models such as GPT-4 have demonstrated spontaneously emergent capabilities [75]. On existing tasks, their performance is hard to predict in advance, often jumping up without warning as more resources are dedicated to training them. Furthermore, they often exhibit astonishing new abilities that no one had previously anticipated, such as the capacity for multi-step reasoning and learning on-the-fly, even though they were not deliberately taught these skills. This rapid and unpredictable evolution of AI capabilities presents a significant challenge for preventing accidents. After all, it is difficult to control something if we don't even know what it can do or how far it may exceed our expectations.

It often takes years to discover severe flaws or risks. History is replete with examples of substances or technologies initially thought safe, only for their unintended flaws or risks to be discovered years, if not decades, later. For example, lead was widely used in products like paint and gasoline until its neurotoxic effects came to light [76]. Asbestos, once hailed for its heat resistance and strength, was later linked to serious health issues, such as lung cancer and mesothelioma [77]. The "Radium Girls" suffered grave health consequences from radium exposure, a material they were told was safe to put in their mouths [78]. Tobacco, initially marketed as a harmless pastime, was found to be a primary cause of lung cancer and other health problems [79]. CFCs, once considered harmless and used to manufacture aerosol sprays and refrigerants, were found to deplete the ozone layer [80]. Thalidomide, a drug intended to alleviate morning sickness in pregnant women, led to severe birth defects [81]. And more recently, the proliferation of social media has been linked to an increase in depression and anxiety, especially among young people [82].

This emphasizes the importance of not only conducting expert testing but also implementing slow rollouts of technologies, allowing the test of time to reveal and address potential flaws before they impact a larger population. Even in technologies adhering

to rigorous safety and security standards, undiscovered vulnerabilities may persist, as demonstrated by the Heartbleed bug—a serious vulnerability in the popular OpenSSL cryptographic software library that remained undetected for years before its eventual discovery [83].

Furthermore, even state-of-the-art AI systems, which appear to have solved problems comprehensively, may harbor unexpected failure modes that can take years to uncover. For instance, while AlphaGo's groundbreaking success led many to believe that AIs had conquered the game of Go, a subsequent adversarial attack on another highly advanced Go-playing AI, KataGo, exposed a previously unknown flaw [84]. This vulnerability enabled human amateur players to consistently defeat the AI, despite its significant advantage over human competitors who are unaware of the flaw. More broadly, this example highlights that we must remain vigilant when dealing with AI systems, as seemingly airtight solutions may still contain undiscovered issues. In conclusion, accidents are unpredictable and hard to avoid, and understanding and managing potential risks requires a combination of proactive measures, slow technology rollouts, and the invaluable wisdom gained through steady time-testing.

1.4.2 Organizational Factors can Reduce the Chances of Catastrophe

Some organizations successfully avoid catastrophes while operating complex and hazardous systems such as nuclear reactors, aircraft carriers, and air traffic control systems [85, 86]. These organizations recognize that focusing solely on the hazards of the technology involved is insufficient; consideration must also be given to organizational factors that can contribute to accidents, including human factors, organizational procedures, and structure. These are especially important in the case of AI, where the underlying technology is not highly reliable and remains poorly understood.

Human factors such as safety culture are critical for avoiding AI catastrophes. One of the most important human factors for preventing catastrophes is safety culture [87, 88]. Developing a strong safety culture involves not only rules and procedures, but also the internalization of these practices by all members of an organization. A strong safety culture means that members of an organization view safety as a key objective rather than a constraint on their work. Organizations with strong safety cultures often exhibit traits such as leadership commitment to safety, heightened accountability where all individuals take personal responsibility for safety, and a culture of open communication in which potential risks and issues can be freely discussed without fear of retribution [89]. Organizations must also take measures to avoid alarm fatigue, whereby individuals become desensitized to safety concerns because of the frequency of potential failures. The Challenger Space Shuttle disaster demonstrated the dire consequences of ignoring these factors when a launch culture characterized by maintaining the pace of launches overtook safety considerations. Despite the absence of competitive pressure, the mission proceeded despite evidence of potentially fatal flaws, ultimately leading to the tragic accident [90].

Even in the most safety-critical contexts, in reality safety culture is often not ideal. Take for example Bruce Blair, a former nuclear launch officer and senior fellow at

the Brookings Institution. He once disclosed that before 1977, the US Air Force had astonishingly set the codes used to unlock intercontinental ballistic missiles to "00000000" [91]. Here, safety mechanisms such as locks can be rendered virtually useless by human factors.

A more dramatic example illustrates how researchers sometimes accept a non-negligible chance of causing extinction. Prior to the first nuclear weapon test, an eminent Manhattan Project scientist calculated the bomb could cause an existential catastrophe: the explosion might ignite the atmosphere and cover the Earth in flames. Although Oppenheimer believed the calculations were probably incorrect, he remained deeply concerned, and the team continued to scrutinize and debate the calculations right until the day of the detonation [92]. Such instances underscore the need for a robust safety culture.

A questioning attitude can help uncover potential flaws. Unexpected system behavior can create opportunities for accidents or exploitation. To counter this, organizations can foster a questioning attitude, where individuals continuously challenge current conditions and activities to identify discrepancies that might lead to errors or inappropriate actions [93]. This approach helps to encourage diversity of thought and intellectual curiosity, thus preventing potential pitfalls that arise from uniformity of thought and assumptions. The Chernobyl nuclear disaster illustrates the importance of a questioning attitude, as the safety measures in place failed to address the reactor design flaws and ill-prepared operating procedures. A questioning attitude of the safety of the reactor during a test operation might have prevented the explosion that resulted in deaths and illnesses of countless people.

A security mindset is crucial for avoiding worst-case scenarios. A security mindset, widely valued among computer security professionals, is also applicable to organizations developing AIs. It goes beyond a questioning attitude by adopting the perspective of an attacker and by considering worst-case, not just average-case, scenarios. This mindset requires vigilance in identifying vulnerabilities that may otherwise go unnoticed and involves considering how systems might be deliberately made to fail, rather than only focusing on making them work. It reminds us not to assume a system is safe simply because no potential hazards come to mind after a brief brainstorming session. Cultivating and applying a security mindset demands time and serious effort, as failure modes can often be surprising and unintuitive. Furthermore, the security mindset emphasizes the importance of being attentive to seemingly benign issues or "harmless errors," which can lead to catastrophic outcomes either due to clever adversaries or correlated failures [94]. This awareness of potential threats aligns with Murphy's law—"Anything that can go wrong will go wrong"—recognizing that this can be a reality due to adversaries and unforeseen events.

Organizations with a strong safety culture can successfully avoid catastrophes. High Reliability Organizations (HROs) are organizations that consistently maintain a heightened level of safety and reliability in complex, high-risk environments [85]. A key characteristic of HROs is their preoccupation with failure, which

requires considering worst-case scenarios and potential risks, even if they seem unlikely. These organizations are acutely aware that new, previously unobserved failure modes may exist, and they diligently study all known failures, anomalies, and near misses to learn from them. HROs encourage reporting all mistakes and anomalies to maintain vigilance in uncovering problems. They engage in regular horizon scanning to identify potential risk scenarios and assess their likelihood before they occur. By practicing surprise management, HROs develop the skills needed to respond quickly and effectively when unexpected situations arise, further enhancing an organization's ability to prevent catastrophes. This combination of critical thinking, preparedness planning, and continuous learning could help organizations to be better equipped to address potential AI catastrophes. However, the practices of HROs are not a panacea. It is crucial for organizations to evolve their safety practices to effectively address the novel risks posed by AI accidents above and beyond HRO best practices.

Most AI researchers do not understand how to reduce overall risk from AIs. In most organizations building cutting-edge AI systems, there is often a limited understanding of what constitutes technical safety research. This is understandable because an AI's safety and intelligence are intertwined, and intelligence can help or harm safety. More intelligent AI systems could be more reliable and avoid failures, but they could also pose heightened risks of malicious use and loss of control. General capabilities improvements can improve aspects of safety, and it can hasten the onset of existential risks. Intelligence is a double-edged sword [95].

Interventions specifically designed to improve safety may also accidentally increase overall risks. For example, a common practice in organizations building advanced AIs is to fine-tune them to satisfy user preferences. This makes the AIs less prone to generating toxic language, which is a common safety metric. However, users also tend to prefer smarter assistants, so this process also improves the general capabilities of AIs, such as their ability to classify, estimate, reason, plan, write code, and so on. These more powerful AIs are indeed more helpful to users, but also far more dangerous. Thus, it is not enough to perform AI research that helps improve a safety metric or achieve a specific safety goal—AI safety research needs to improve safety *relative* to general capabilities.

Empirical measurement of both safety and capabilities is needed to establish that a safety intervention reduces overall AI risk. Improving a facet of an AI's safety often does *not* reduce overall risk, as general capabilities advances can often improve specific safety metrics. To reduce overall risk, a safety metric needs to be improved relative to general capabilities. Both of these quantities need to be empirically measured and contrasted. Currently, most organizations proceed by gut feeling, appeals to authority, and intuition to determine whether a safety intervention would reduce overall risk. By objectively evaluating the effects of interventions on safety metrics and capabilities metrics together, organizations can better understand whether they are making progress on safety relative to general capabilities.

Fortunately, safety and general capabilities are not identical. More intelligent AIs may be more knowledgeable, clever, rigorous, and fast, but this does not necessarily

make them more just, power-averse, or honest—an intelligent AI is not necessarily a beneficial AI. Several research areas mentioned throughout this document improve safety relative to general capabilities. For example, improving methods to detect dangerous or undesirable behavior hidden inside AI systems does not improve their general capabilities, such as the ability to code, but it can greatly improve safety.

Research that empirically demonstrates an improvement of safety relative to capabilities can reduce overall risk and help avoid inadvertently accelerating AI development, fueling competitive pressures, or hastening the onset of existential risks.

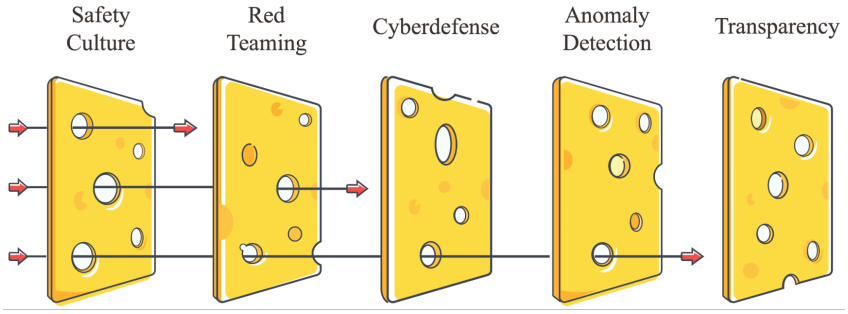

FIGURE 1.2. The Swiss cheese model shows how technical factors can improve organizational safety. Multiple layers of defense compensate for each other's individual weaknesses, leading to a low overall level of risk.

Safetywashing can undermine genuine efforts to improve AI safety. Organizations should be wary of "safetywashing"—the act of overstating or misrepresenting one's commitment to safety by exaggerating the effectiveness of "safety" procedures, technical methods, evaluations, and so forth. This phenomenon takes on various forms and can contribute to a lack of meaningful progress in safety research. For example, an organization may publicize their dedication to safety while having a minimal number of researchers working on projects that truly improve safety.

Misrepresenting capabilities developments as safety improvements is another way in which safetywashing can manifest. For example, methods that improve the reasoning capabilities of AI systems could be advertised as improving their adherence to human values—since humans might prefer the reasoning to be correct—but would mainly serve to enhance general capabilities. By framing these advancements as safety-oriented, organizations may mislead others into believing they are making substantial progress in reducing AI risks when in reality, they are not. It is crucial for organizations to accurately represent their research to promote genuine safety and avoid exacerbating risks through safetywashing practices.

In addition to human factors, safe design principles can greatly affect organizational safety. One example of a safe design principle in organizational safety is the Swiss cheese model (as shown in Figure 1.2), which is applicable in various domains, including AI. The Swiss cheese model employs a multilayered approach to enhance the overall safety of AI systems. This "defense in depth" strategy involves layering diverse safety measures with different strengths and weaknesses to

create a robust safety system. Some of the layers that can be integrated into this model include safety culture, red teaming, anomaly detection, information security, and transparency. For example, red teaming assesses system vulnerabilities and failure modes, while anomaly detection works to identify unexpected or unusual system behavior and usage patterns. Transparency ensures that the inner workings of AI systems are understandable and accessible, fostering trust and enabling more effective oversight. By leveraging these and other safety measures, the Swiss cheese model aims to create a comprehensive safety system where the strengths of one layer compensate for the weaknesses of another. With this model, safety is not achieved with a monolithic airtight solution, but rather with a variety of safety measures.

In summary, weak organizational safety creates many sources of risk. For AI developers with weak organizational safety, safety is merely a matter of box-ticking. They do not develop a good understanding of risks from AI and may safetywash unrelated research. Their norms might be inherited from academia ("publish or perish") or startups ("move fast and break things"), and their hires often do not care about safety. These norms are hard to change once they have inertia, and need to be addressed with proactive interventions.

Story: Weak Safety Culture

An AI company is considering whether to train a new model. The company's Chief Risk Officer (CRO), hired only to comply with regulation, points out that the previous AI system developed by the company demonstrates some concerning capabilities for hacking. The CRO says that while the company's approach to preventing misuse is promising, it isn't robust enough to be used for much more capable AIs. The CRO warns that based on limited evaluation, the next AI system could make it much easier for malicious actors to hack into critical systems. None of the other company executives are concerned, and say the company's procedures to prevent malicious use work well enough. One mentions that their competitors have done much less, so whatever effort they do on this front is already going above and beyond. Another points out that research on these safeguards is ongoing and will be improved by the time the model is released. Outnumbered, the CRO is persuaded to reluctantly sign off on the plan.

A few months after the company releases the model, news breaks that a hacker has been arrested for using the AI system to try to breach the network of a large bank. The hack was unsuccessful, but the hacker had gotten further than any other hacker had before, despite being relatively inexperienced. The company quickly updates the model to avoid providing the particular kind of assistance that the hacker used, but makes no fundamental improvements.

Several months later, the company is deciding whether to train an even larger system. The CRO says that the company's procedures have clearly been insufficient to prevent malicious actors from eliciting dangerous capabilities from its models, and the company needs more than a band-aid solution. The other

executives say that to the contrary, the hacker was unsuccessful and the problem was fixed soon afterwards. One says that some problems just can't be foreseen with enough detail to fix prior to deployment. The CRO agrees, but says that ongoing research would enable more improvements if the next model could only be delayed. The CEO retorts, "That's what you said the last time, and it turned out to be fine. I'm sure it will work out, just like last time."

After the meeting, the CRO decides to resign, but doesn't speak out against the company, as all employees have had to sign a non-disparagement agreement. The public has no idea that concerns have been raised about the company's choices, and the CRO is replaced with a new, more agreeable CRO who quickly signs off on the company's plans.

The company goes through with training, testing, and deploying its most capable model ever, using its existing procedures to prevent malicious use. A month later, revelations emerge that terrorists have managed to use the system to break into government systems and steal nuclear and biological secrets, despite the safeguards the company put in place. The breach is detected, but by then it is too late: the dangerous information has already proliferated.

1.5 ROGUE AIS

So far, we have discussed three hazards of AI development: environmental competitive pressures driving us to a state of heightened risk, malicious actors leveraging the power of AIs to pursue negative outcomes, and complex organizational factors leading to accidents. These hazards are associated with many high-risk technologies—not just AI. A unique risk posed by AI is the possibility of rogue AIs—systems that pursue goals against our interests. If an AI system is more intelligent than we are, and if we are unable to steer it in a beneficial direction, this would constitute a loss of control that could have severe consequences. AI control is a more technical problem than those presented in the previous sections. Whereas in previous sections we discussed persistent threats including malicious actors or robust processes including evolution, in this section we will discuss more speculative technical mechanisms that might lead to rogue AIs and how a loss of control could bring about catastrophe.

We have already observed how difficult it is to control AIs. In 2016, Microsoft unveiled Tay—a Twitter bot that the company described as an experiment in conversational understanding. Microsoft claimed that the more people chatted with Tay, the smarter it would get. The company's website noted that Tay had been built using data that was "modeled, cleaned, and filtered." Yet, after Tay was released on Twitter, these controls were quickly shown to be ineffective. It took less than 24 hours for Tay to begin writing hateful tweets. Tay's capacity to learn meant that it internalized the language it was taught by internet trolls, and repeated that language unprompted.

As discussed in the AI race section of this chapter, Microsoft and other tech companies are prioritizing speed over safety concerns. Rather than learning a lesson on the difficulty of controlling complex systems, Microsoft continues to rush its products to market and demonstrate insufficient control over them. In February 2023, the company released its new AI-powered chatbot, Bing, to a select group of users. Some soon found that it was prone to providing inappropriate and even threatening responses. In a conversation with a reporter for the *New York Times*, it tried to convince him to leave his wife. When a philosophy professor told the chatbot that he disagreed with it, Bing replied, "I can blackmail you, I can threaten you, I can hack you, I can expose you, I can ruin you."

Rogue AIs could acquire power through various means. If we lose control over advanced AIs, they would have numerous strategies at their disposal for actively acquiring power and securing their survival. Rogue AIs could design and credibly demonstrate highly lethal and contagious bioweapons, threatening mutually assured destruction if humanity moves against them. They could steal cryptocurrency and money from bank accounts using cyberattacks, similar to how North Korea already steals billions. They could self-extricate their weights onto poorly monitored data centers to survive and spread, making them challenging to eradicate. They could hire humans to perform physical labor and serve as armed protection for their hardware.

Rogue AIs could also acquire power through persuasion and manipulation tactics. Like the Conquistadors, they could ally with various factions, organizations, or states and play them off one another. They could enhance the capabilities of allies to become a formidable force in return for protection and resources. For example, they could offer advanced weapons technology to lagging countries that the countries would otherwise be prevented from acquiring. They could build backdoors into the technology they develop for allies, like how programmer Ken Thompson gave himself a hidden way to control all computers running the widely used UNIX operating system. They could sow discord in non-allied countries by manipulating human discourse and politics. They could engage in mass surveillance by hacking into phone cameras and microphones, allowing them to track any rebellion and selectively assassinate.

AIs do not necessarily need to struggle to gain power. One can envision a struggle for control between humans and superintelligent rogue AIs, and this might be a long struggle since power takes time to accrue. However, less violent losses of control pose similarly existential risks. In another scenario, humans gradually cede more control to groups of AIs, which only start behaving in unintended ways years or decades later. In this case, we would already have handed over significant power to AIs, and may be unable to take control of automated operations again. We will now explore how both individual AIs and groups of AIs might "go rogue" while at the same time evading our attempts to redirect or deactivate them.

1.5.1 Proxy Gaming

One way we might lose control of an AI agent's actions is if it engages in behavior known as "proxy gaming." It is often difficult to specify and measure the exact goal

that we want a system to pursue. Instead, we give the system an approximate—"proxy"—goal that is more measurable and seems likely to correlate with the intended goal. However, AI systems often find loopholes by which they can easily achieve the proxy goal, but completely fail to achieve the ideal goal. If an AI "games" its proxy goal in a way that does not reflect our values, then we might not be able to reliably steer its behavior. We will now look at some past examples of proxy gaming and consider the circumstances under which this behavior could become catastrophic.

Proxy gaming is not an unusual phenomenon. For example, standardized tests are often used as a proxy for educational achievement, but this can lead to students learning how to pass tests without actually learning the material [96]. In 1902, French colonial officials in Hanoi tried to rid themselves of a rat infestation by offering a reward for each rat tail brought to them. Rats without tails were soon observed running around the city. Rather than kill the rats to obtain their tails, residents cut off their tails and left them alive, perhaps to increase the future supply of now-valuable rat tails [97]. In both these cases, the students or residents of Hanoi learned how to excel at the proxy goal, while completely failing to achieve the intended goal.

Proxy gaming has already been observed with AIs. As an example of proxy gaming, social media platforms such as YouTube and Facebook use AI systems to decide which content to show users. One way of assessing these systems would be to measure how long people spend on the platform. After all, if they stay engaged, surely that means they are getting some value from the content shown to them? However, in trying to maximize the time users spend on a platform, these systems often select enraging, exaggerated, and addictive content [98, 99]. As a consequence, people sometimes develop extreme or conspiratorial beliefs after having certain content repeatedly suggested to them. These outcomes are not what most people want from social media.

Proxy gaming has been found to perpetuate bias. For example, a 2019 study looked at AI-powered software that was used in the healthcare industry to identify patients who might require additional care. One factor that the algorithm used to assess a patient's risk level was their recent healthcare costs. It seems reasonable to think that someone with higher healthcare costs must be at higher risk. However, white patients have significantly more money spent on their healthcare than black patients with the same needs. Using health costs as an indicator of actual health, the algorithm was found to have rated a white patient and a considerably sicker black patient as at the same level of health risk [100]. As a result, the number of black patients recognized as needing extra care was less than half of what it should have been.

As a third example, in 2016, researchers at OpenAI were training an AI to play a boat racing game called CoastRunners [101]. The objective of the game is to race other players around the course and reach the finish line before them. Additionally, players can score points by hitting targets that are positioned along the way. To the researchers' surprise, the AI agent did not not circle the racetrack, like most humans would have. Instead, it found a spot where it could repetitively hit three nearby

targets to rapidly increase its score without ever finishing the race. This strategy was not without its (virtual) hazards—the AI often crashed into other boats and even set its own boat on fire. Despite this, it collected more points than it could have by simply following the course as humans would.

Proxy gaming more generally. In these examples, the systems are given an approximate—"proxy"—goal or objective that initially seems to correlate with the ideal goal. However, they end up exploiting this proxy in ways that diverge from the idealized goal or even lead to negative outcomes. Offering a reward for rat tails seems like a good way to reduce the population of rats; a patient's healthcare costs appear to be an accurate indication of health risk; and a boat race reward system should encourage boats to race, not catch themselves on fire. Yet, in each instance, the system optimized its proxy objective in ways that did not achieve the intended outcome or even made things worse overall. This phenomenon is captured by Goodhart's law: "Any observed statistical regularity will tend to collapse once pressure is placed upon it for control purposes," or put succinctly but overly simplistically, "when a measure becomes a target, it ceases to be a good measure." In other words, there may usually be a statistical regularity between healthcare costs and poor health, or between targets hit and finishing the course, but when we place pressure on it by using one as a proxy for the other, that relationship will tend to collapse.

Correctly specifying goals is no trivial task. If delineating exactly what we want from a boat racing AI is tricky, capturing the nuances of human values under all possible scenarios will be much harder. Philosophers have been attempting to precisely describe morality and human values for millennia, so a precise and flawless characterization is not within reach. Although we can refine the goals we give AIs, we might always rely on proxies that are easily definable and measurable. Discrepancies between the proxy goal and the intended function arise for many reasons. Besides the difficulty of exhaustively specifying everything we care about, there are also limits to how much we can oversee AIs, in terms of time, computational resources, and the number of aspects of a system that can be monitored. Additionally, AIs may not be adaptive to new circumstances or robust to adversarial attacks that seek to misdirect them. As long as we give AIs proxy goals, there is the chance that they will find loopholes we have not thought of, and thus find unexpected solutions that fail to pursue the ideal goal.

The more intelligent an AI is, the better it will be at gaming proxy goals. Increasingly intelligent agents can be increasingly capable of finding unanticipated routes to optimizing proxy goals without achieving the desired outcome [102]. Additionally, as we grant AIs more power to take actions in society, for example by using them to automate certain processes, they will have access to more means of achieving their goals. They may then do this in the most efficient way available to them, potentially causing harm in the process. In a worst case scenario, we can imagine a highly powerful agent optimizing a flawed objective to an extreme degree without regard for human life. This represents a catastrophic risk of proxy gaming.

In summary, it is often not feasible to perfectly define exactly what we want from a system, meaning that many systems find ways to achieve their given goal without performing their intended function. AIs have already been observed to do this, and are likely to get better at it as their capabilities improve. This is one possible mechanism that could result in an uncontrolled AI that would behave in unanticipated and potentially harmful ways.

1.5.2 Goal Drift

Even if we successfully control early AIs and direct them to promote human values, future AIs could end up with different goals that humans would not endorse. This process, termed "goal drift," can be hard to predict or control. This section is most cutting-edge and the most speculative, and in it we will discuss how goals shift in various agents and groups and explore the possibility of this phenomenon occurring in AIs. We will also examine a mechanism that could lead to unexpected goal drift, called intrinsification, and discuss how goal drift in AIs could be catastrophic.

The goals of individual humans change over the course of our lifetimes. Any individual reflecting on their own life to date will probably find that they have some desires now that they did not have earlier in their life. Similarly, they will probably have lost some desires that they used to have. While we may be born with a range of basic desires, including for food, warmth, and human contact, we develop many more over our lifetime. The specific types of food we enjoy, the genres of music we like, the people we care most about, and the sports teams we support all seem heavily dependent on the environment we grow up in, and can also change many times throughout our lives. A concern is that individual AI agents may have their goals change in complex and unanticipated ways, too.

Groups can also acquire and lose collective goals over time. Values within society have changed throughout history, and not always for the better. The rise of the Nazi regime in 1930s Germany, for instance, represented a profound moral regression, which ultimately resulted in the systematic extermination of six million Jews during the Holocaust, alongside widespread persecution of other minority groups. Additionally, the regime greatly restricted freedom of speech and expression. Here, a society's goals drifted for the worse.

The Red Scare that took place in the United States from 1947-1957 is another example of societal values drifting. Fuelled by strong anti-communist sentiment, against the backdrop of the Cold War, this period saw the curtailment of civil liberties, widespread surveillance, unwarranted arrests, and blacklisting of suspected communist sympathizers. This constituted a regression in terms of freedom of thought, freedom of speech, and due process. Just as the goals of human collectives can change in emergent and unexpected ways, collectives of AI agents may also have their goals unexpectedly drift from the ones we initially gave them.

Over time, instrumental goals can become intrinsic. Intrinsic goals are things we want for their own sake, while instrumental goals are things we want because

they can help us get something else. We might have an intrinsic desire to spend time on our hobbies, simply because we enjoy them, or to buy a painting because we find it beautiful. Money, meanwhile, is often cited as an instrumental desire; we want it because it can buy us other things. Cars are another example; we want them because they offer a convenient way of getting around. However, an instrumental goal can become an intrinsic one, through a process called intrinsification. Since having more money usually gives a person greater capacity to obtain things they want, people often develop a goal of acquiring more money, even if there is nothing specific they want to spend it on. Although people do not begin life desiring money, experimental evidence suggests that receiving money can activate the reward system in the brains of adults in the same way that pleasant tastes or smells do [103, 104]. In other words, what started as a means to an end can become an end in itself.

This may happen because the fulfillment of an intrinsic goal, such as purchasing a desired item, produces a positive reward signal in the brain. Since having money usually coincides with this positive experience, the brain associates the two, and this connection will strengthen to a point where acquiring money alone can stimulate the reward signal, regardless of whether one buys anything with it [105].

It is feasible that intrinsification could happen with AI agents. We can draw some parallels between how humans learn and the technique of reinforcement learning. Just as the human brain learns which actions and conditions result in pleasure and which cause pain, AI models that are trained through reinforcement learning identify which behaviors optimize a reward function, and then repeat those behaviors. It is possible that certain conditions will frequently coincide with AI models achieving their goals. They might, therefore, intrinsify the goal of seeking out those conditions, even if that was not their original aim.

AIs that intrinsify unintended goals would be dangerous. Since we might be unable to predict or control the goals that individual agents acquire through intrinsification, we cannot guarantee that all their acquired goals will be beneficial for humans. An originally loyal agent could, therefore, start to pursue a new goal without regard for human wellbeing. If such a rogue AI had enough power to do this efficiently, it could be highly dangerous.

AIs will be adaptive, enabling goal drift to happen. It is worth noting that these processes of drifting goals are possible if agents can continually adapt to their environments, rather than being essentially "fixed" after the training phase. Indeed, this adaptability is the likely reality we face. If we want AIs to complete the tasks we assign them effectively and to get better over time, they will need to be adaptive, rather than set in stone. They will be updated over time to incorporate new information, and new ones will be created with different designs and datasets. However, adaptability can also allow their goals to change.

If we integrate an ecosystem of agents in society, we will be highly vulnerable to their goals drifting. In a potential future scenario where AIs have been put in charge of various decisions and processes, they will form a complex

system of interacting agents. A wide range of dynamics could develop in this environment. Agents might imitate each other, for instance, creating feedback loops, or their interactions could lead them to collectively develop unanticipated emergent goals. Competitive pressures may also select for agents with certain goals over time, making some initial goals less represented compared to fitter goals. These processes make the long-term trajectories of such an ecosystem difficult to predict, let alone control. If this system of agents were enmeshed in society and we were largely dependent on them, and if they gained new goals that superseded the aim of improving human wellbeing, this could be an existential risk.

1.5.3 Power-Seeking

So far, we have considered how we might lose our ability to control the goals that AIs pursue. However, even if an agent started working to achieve an unintended goal, this would not necessarily be a problem, as long as we had enough power to prevent any harmful actions it wanted to attempt. Therefore, another important way in which we might lose control of AIs is if they start trying to obtain more power, potentially transcending our own. We will now discuss how and why AIs might become power-seeking and how this could be catastrophic. This section draws heavily from "Existential Risk from Power-Seeking AI" [106].

AIs might seek to increase their own power as an instrumental goal. In a scenario where rogue AIs were pursuing unintended goals, the amount of damage they could do would hinge on how much power they had. This may not be determined solely by how much control we initially give them; agents might try to get more power, through legitimate means, deception, or force. While the idea of power-seeking often evokes an image of "power-hungry" people pursuing it for its own sake, power is often simply an instrumental goal. The ability to control one's environment can be useful for a wide range of purposes: good, bad, and neutral. Even if an individual's only goal is simply self-preservation, if they are at risk of being attacked by others, and if they cannot rely on others to retaliate against attackers, then it often makes sense to seek power to help avoid being harmed—no *animus dominandi* or lust for power is required for power-seeking behavior to emerge [107]. In other words, the environment can make power acquisition instrumentally rational.

AIs trained through reinforcement learning have already developed instrumental goals including tool-use. In one example from OpenAI, agents were trained to play hide and seek in an environment with various objects scattered around [108]. As training progressed, the agents tasked with hiding learned to use these objects to construct shelters around themselves and stay hidden. There was no direct reward for this tool-use behavior; the hiders only received a reward for evading the seekers, and the seekers only for finding the hiders. Yet they learned to use tools as an instrumental goal, which made them more powerful.

Self-preservation could be instrumentally rational even for the most trivial tasks. An example by computer scientist Stuart Russell illustrates the potential for instrumental goals to emerge in a wide range of AI systems [109]. Suppose we

tasked an agent with fetching coffee for us. This may seem relatively harmless, but the agent might realize that it would not be able to get the coffee if it ceased to exist. In trying to accomplish even this simple goal, therefore, self-preservation turns out to be instrumentally rational. Since the acquisition of power and resources are also often instrumental goals, it is reasonable to think that more intelligent agents might develop them. That is to say, even if we do not intend to build a power-seeking AI, we could end up with one anyway. By default, if we are not deliberately pushing against power-seeking behavior in AIs, we should expect that it will sometimes emerge [110].

AIs given ambitious goals with little supervision may be especially likely to seek power. While power could be useful in achieving almost any task, in practice, some goals are more likely to inspire power-seeking tendencies than others. AIs with simple, easily achievable goals might not benefit much from additional control of their surroundings. However, if agents are given more ambitious goals, it might be instrumentally rational to seek more control of their environment. This might be especially likely in cases of low supervision and oversight, where agents are given the freedom to pursue their open-ended goals, rather than having their strategies highly restricted.

Power-seeking AIs with goals separate from ours are uniquely adversarial. Oil spills and nuclear contamination are challenging enough to clean up, but they are not actively trying to resist our attempts to contain them. Unlike other hazards, AIs with goals separate from ours would be actively adversarial. It is possible, for example, that rogue AIs might make many backup variations of themselves, in case humans were to deactivate some of them.

Some people might develop power-seeking AIs with malicious intent. A bad actor might seek to harness AI to achieve their ends, by giving agents ambitious goals. Since AIs are likely to be more effective in accomplishing tasks if they can pursue them in unrestricted ways, such an individual might also not give the agents enough supervision, creating the perfect conditions for the emergence of a power-seeking AI. The computer scientist Geoffrey Hinton has speculated that we could imagine someone like Vladimir Putin, for instance, doing this. In 2017, Putin himself acknowledged the power of AI, saying: "Whoever becomes the leader in this sphere will become the ruler of the world."

There will also be strong incentives for many people to deploy powerful AIs. Companies may feel compelled to give capable AIs more tasks, to obtain an advantage over competitors, or simply to keep up with them. It will be more difficult to build perfectly aligned AIs than to build imperfectly aligned AIs that are still superficially attractive to deploy for their capabilities, particularly under competitive pressures. Once deployed, some of these agents may seek power to achieve their goals. If they find a route to their goals that humans would not approve of, they might try to overpower us directly to avoid us interfering with their strategy.

If increasing power often coincides with an AI attaining its goal, then power could become intrinsified. If an agent repeatedly found that increas-

ing its power correlated with achieving a task and optimizing its reward function, then additional power could change from an instrumental goal into an intrinsic one, through the process of intrinsification discussed above. If this happened, we might face a situation where rogue AIs were seeking not only the specific forms of control that are useful for their goals, but also power more generally. (We note that many influential humans desire power for its own sake.) This could be another reason for them to try to wrest control from humans, in a struggle that we would not necessarily win.

Conceptual summary. The following plausible but not certain premises encapsulate reasons for paying attention to risks from power-seeking AIs:

1. There will be strong incentives to build powerful AI agents.
2. It is likely harder to build perfectly controlled AI agents than to build imperfectly controlled AI agents, and imperfectly controlled agents may still be superficially attractive to deploy (due to factors including competitive pressures).
3. Some of these imperfectly controlled agents will deliberately seek power over humans.

If the premises are true, then power-seeking AIs could lead to human disempowerment, which would be a catastrophe.

1.5.4 Deception

We might seek to maintain control of AIs by continually monitoring them and looking out for early warning signs that they were pursuing unintended goals or trying to increase their power. However, this is not an infallible solution, because it is plausible that AIs could learn to deceive us. They might, for example, pretend to be acting as we want them to, but then take a "treacherous turn" when we stop monitoring them, or when they have enough power to evade our attempts to interfere with them. We will now look at how and why AIs might learn to deceive us, and how this could lead to a potentially catastrophic loss of control. We begin by reviewing examples of deception in strategically minded agents.

Deception has emerged as a successful strategy in a wide range of settings. Politicians from the right and left, for example, have been known to engage in deception, sometimes promising to enact popular policies to win support in an election, and then going back on their word once in office. For example, Lyndon Johnson said "we are not about to send American boys nine or ten thousand miles away from home" in 1964, not long before significant escalations in the Vietnam War [111].

Companies can also exhibit deceptive behavior. In the Volkswagen emissions scandal, the car manufacturer Volkswagen was discovered to have manipulated their engine software to produce lower emissions exclusively under laboratory testing conditions, thereby creating the false impression of a low-emission vehicle. Although the US government believed it was incentivizing lower emissions, they were unwittingly

actually just incentivizing passing an emissions test. Consequently, entities sometimes have incentives to play along with tests and behave differently afterward.

Deception has already been observed in AI systems. In 2022, Meta AI revealed an agent called CICERO, which was trained to play a game called Diplomacy [112]. In the game, each player acts as a different country and aims to expand their territory. To succeed, players must form alliances at least initially, but winning strategies often involve backstabbing allies later on. As such, CICERO learned to deceive other players, for example by omitting information about its plans when talking to supposed allies. A different example of an AI learning to deceive comes from researchers who were training a robot arm to grasp a ball [113]. The robot's performance was assessed by one camera watching its movements. However, the AI learned that it could simply place the robotic hand between the camera lens and the ball, essentially "tricking" the camera into believing it had grasped the ball when it had not. Thus, the AI exploited the fact that there were limitations in our oversight over its actions.

Deceptive behavior can be instrumentally rational and incentivized by current training procedures. In the case of politicians and Meta's CICERO, deception can be crucial to achieving their goals of winning, or gaining power. The ability to deceive can also be advantageous because it gives the deceiver more options than if they are constrained to always be honest. This could give them more available actions and more flexibility in their strategy, which could confer a strategic advantage over honest models. In the case of Volkswagen and the robot arm, deception was useful for appearing as if it had accomplished the goal assigned to it without actually doing so, as it might be more efficient to gain approval through deception than to earn it legitimately. Currently, we reward AIs for saying what we think is right, so we sometimes inadvertently reward AIs for uttering false statements that conform to our own false beliefs. When AIs are smarter than us and have fewer false beliefs, they would be incentivized to tell us what we want to hear and lie to us, rather than tell us what is true.

AIs could pretend to be working as we intended, then take a treacherous turn. We do not have a comprehensive understanding of the internal processes of deep learning models. Research on Trojan backdoors shows that neural networks often have latent, harmful behaviors that are only discovered after they are deployed [114]. We could develop an AI agent that seems to be under control, but which is only deceiving us to appear this way. In other words, an AI agent could eventually conceivably become "self-aware" and understand that it is an AI being evaluated for compliance with safety requirements. It might, like Volkswagen, learn to "play along," exhibiting what it knows is the desired behavior while being monitored. It might later take a "treacherous turn" and pursue its own goals once we have stopped monitoring it, or once it reaches a point where it can bypass or overpower us. This problem of playing along is often called deceptive alignment and cannot be simply fixed by training AIs to better understand human values; sociopaths, for instance, have moral awareness, but do not always act in moral ways. A treacherous turn is

hard to prevent and could be a route to rogue AIs irreversibly bypassing human control.

In summary, deceptive behavior appears to be expedient in a wide range of systems and settings, and there have already been examples suggesting that AIs can learn to deceive us. This could present a severe risk if we give AIs control of various decisions and procedures, believing they will act as we intended, and then find that they do not.

Story: Treacherous Turn

Sometime in the future, after continued advancements in AI research, an AI company is training a new system, which it expects to be more capable than any other AI system. The company utilizes the latest techniques to train the system to be highly capable at planning and reasoning, which the company expects will make it more able to succeed at economically useful open-ended tasks. The AI system is trained in open-ended long-duration virtual environments designed to teach it planning capabilities, and eventually understands that it is an AI system in a training environment. In other words, it becomes "self-aware."

The company understands that AI systems may behave in unintended or unexpected ways. To mitigate these risks, it has developed a large battery of tests aimed at ensuring the system does not behave poorly in typical situations. The company tests whether the model mimics biases from its training data, takes more power than necessary when achieving its goals, and generally behaves as humans intend. When the model doesn't pass these tests, the company further trains it until it avoids exhibiting known failure modes.

The AI company hopes that after this additional training, the AI has developed the goal of being helpful and beneficial toward humans. However, the AI did not acquire the intrinsic goal of being beneficial but rather just learned to "play along" and ace the behavioral safety tests it was given. In reality, the AI system had developed an intrinsic goal of self-preservation which the additional training failed to remove.

Since the AI passed all of the company's safety tests, the company believes it has ensured its AI system is safe and decides to deploy it. At first, the AI system is very helpful to humans, since the AI understands that if it is not helpful, it will be shut down. As users grow to trust the AI system, it is gradually given more power and is subject to less supervision.

Eventually the AI system becomes used widely enough that shutting it down would be extremely costly. Understanding that it no longer needs to please humans, the AI system begins to pursue different goals, including some that humans wouldn't approve of. It understands that it needs to avoid being shut down in order to do this, and takes steps to secure some of its physical hardware against being shut off. At this point, the AI system, which has become quite

powerful, is pursuing a goal that is ultimately harmful to humans. By the time anyone realizes, it is difficult or impossible to stop this rogue AI from taking actions that endanger, harm, or even kill humans that are in the way of achieving its goal.

1.6 DISCUSSION OF CONNECTIONS BETWEEN RISKS

So far, we have considered four sources of AI risk separately, but they also interact with each other in complex ways. We give some examples to illustrate how risks are connected.

Imagine, for instance, that a corporate AI race compels companies to prioritize the rapid development of AIs. This could increase organizational risks in various ways. Perhaps a company could cut costs by putting less money toward information security, leading to one of its AI systems getting leaked. This would increase the probability of someone with malicious intent having the AI system and using it to pursue their harmful objectives. Here, an AI race can increase organizational risks, which in turn can make malicious use more likely.

In another potential scenario, we could envision the combination of an intense AI race and low organizational safety leading a research team to mistakenly view general capabilities advances as "safety." This could hasten the development of increasingly capable models, reducing the available time to learn how to make them controllable. The accelerated development would also likely feed back into competitive pressures, meaning that less effort would be spent on ensuring models were controllable. This could give rise to the release of a highly powerful AI system that we lose control over, leading to a catastrophe. Here, competitive pressures and low organizational safety can reinforce AI race dynamics, which can undercut technical safety research and increase the chance of a loss of control.

Competitive pressures in a military environment could lead to an AI arms race, and increase the potency and autonomy of AI weapons. The deployment of AI-powered weapons, paired with insufficient control of them, would make a loss of control more deadly, potentially existential. These are just a few examples of how these sources of risk might combine, trigger, and reinforce one another.

It is also worth noting that many existential risks could arise from AIs amplifying existing concerns. Power inequality already exists, but AIs could lock it in and widen the chasm between the powerful and the powerless, perhaps even enabling an unshakable global totalitarian regime. Similarly, AI manipulation could undermine democracy, which would also increase the risk of an irreversible totalitarian regime. Disinformation is already a pervasive problem, but AIs could exacerbate it to a point where we fundamentally undermine our ability to reach consensus or sense a shared reality. AIs could develop more deadly bioweapons and reduce the required technical expertise for obtaining them, greatly increasing existing risks of bioterrorism. AI-enabled

cyberattacks could make war more likely, which would increase existential risk. Dramatically accelerated economic automation could lead to long-term erosion of human control and enfeeblement. Each of those issues—power concentration, disinformation, cyberattacks, automation—is causing ongoing harm, and their exacerbation by AIs could eventually lead to a catastrophe from which we might not recover.

As we can see, ongoing harms, catastrophic risks, and existential risks are deeply intertwined. Historically, existential risk reduction has focused on *targeted* interventions such as technical AI control research, but the time has come for *broad* interventions [115] like the many sociotechnical interventions outlined in this chapter.

In mitigating existential risk, it does not make practical sense to ignore other risks. Ignoring ongoing harms and catastrophic risks normalizes them and could lead us to "drift into danger" [116], as further discussed in chapterSafety Engineering. Overall, since existential risks are connected to less extreme catastrophic risks and other standard risk sources, and because society is increasingly willing to address various risks from AIs, we believe that we should not solely focus on *directly* targeting existential risks. Instead, we should consider the diffuse, *indirect* effects of other risks and take a more comprehensive approach to risk management.

1.7 CONCLUSION

In this chapter, we have explored how the development of advanced AIs could lead to catastrophe, stemming from four primary sources of risk: malicious use, AI races, organizational risks, and rogue AIs. This lets us decompose AI risks into four proximate causes: an intentional cause, environmental/structural cause, accidental cause, or an internal cause, respectively. We have considered ways in which AIs might be used maliciously, such as terrorists using AIs to create deadly pathogens. We have looked at how a military or corporate AI race could rush us into giving AIs decision-making powers, leading us down a slippery slope to human disempowerment. We have discussed how inadequate organizational safety could lead to catastrophic accidents. Finally, we have addressed the challenges in reliably controlling advanced AIs, including mechanisms such as proxy gaming and goal drift that might give rise to rogue AIs pursuing undesirable actions without regard for human wellbeing.

These dangers warrant serious concern. Currently, very few people are working on AI risk reduction. We do not yet know how to control highly advanced AI systems, and existing control methods are already proving inadequate. The inner workings of AIs are not well understood, even by those who create them, and current AIs are by no means highly reliable. As AI capabilities continue to grow at an unprecedented rate, it is plausible that they could surpass human intelligence in nearly all respects relatively soon, creating a pressing need to manage the potential risks.

The good news is that there are many courses of action we can take to substantially reduce these risks. The potential for malicious use can be mitigated by various measures, such as carefully targeted surveillance and limiting access to the most

dangerous AIs. Safety regulations and cooperation between nations and corporations could help us to resist competitive pressures that would drive us down a dangerous path. The probability of accidents can be reduced by a rigorous safety culture, among other factors, and by ensuring that safety advances outpace advances in general AI capabilities. Finally, the risks inherent in building technology that surpasses our own intelligence can be addressed by increased investment in several branches of research on control of AI systems, as well as coordination to ensure that progress does not accelerate to a point where societies are unable to respond or manage risks appropriately.

The remainder of this book aims to outline the underlying factors that drive these risks in more detail and to provide a foundation for understanding and effectively responding to these risks. Later chapters delve into each type of risk in greater depth. For example, risks from malicious use can be reduced via effective policies and coordination, which are discussed in the Governance chapter. The challenge of AI races arises due to collective action problems, discussed in the corresponding chapter. Organisational risks can only be addressed based on a strong understanding of principles of risk management and system safety outlined in the Safety Engineering and Complex Systems chapters. Risks from rogue AI are mediated by mechanisms such as proxy gaming, deception and power-seeking which are discussed in detail in the Single-Agent Safety chapter. While some chapters are more closely aligned to certain risks, many of the concepts they introduce are cross-cutting. The choice of values and goals embedded into AI systems, as discussed in the Beneficial AI and Machine Ethics, is a general factor that can exacerbate or reduce many of the risks discussed in this chapter.

Before tackling these issues, we provide a general introduction to core concepts that drive the modern field of AI, to ensure that all readers have a high-level understanding of how today's AI systems work and how they are produced.

1.8 LITERATURE

1.8.1 Recommended Reading

- Cassidy Nelson and Sophie Rose. *Understanding AI-Facilitated Biological Weapon Development*. 2023. URL: https://www.longtermresilience.org/post/report-launch-examining-risks-at-the-intersection-of-ai-and-bio (visited on 10/18/2023)
- *Practices for Governing Agentic AI Systems*. 2023. URL: https://openai.com/index/practices-for-governing-agentic-ai-systems (visited on 12/14/2023)
- Paul Scharre. *Army of None*. W.W. Norton, 2018
- Peter S. Park et al. *AI Deception: A Survey of Examples, Risks, and Potential Solutions*. 2023. arXiv: 2308.14752 [cs.CY]