Safety Not Guaranteed:

International Races for Risky Technologies

Eoghan Stafford, Robert F. Trager, Allan Dafoe

November 2022

Abstract

The great powers appear to be entering an era of heightened competition to master security-relevant technologies in areas such as AI. This is concerning because deploying new technologies can create substantial shared risks, such as inadvertent crisis escalation or uncontrolled proliferation. We analyze a strategic model to determine when states deploy technologies before learning how to minimize their risks. When competitors are moderately adversarial or the technology laggard is not very capable, the laggard does not use a risky technology unless it catches up to the technology leader. By contrast, if competitors are highly adversarial and the laggard is closer to the leader's capability level, the laggard is willing to cut corners to gamble for advantage, so that the shared risk *falls* if the laggard catches up. Further, when competitors are not deploying the riskiest technologies, steps to make those technologies safer will be attenuated or reversed by risk compensation.

Preliminary. Please do not cite or distribute without permission.

We are grateful for outstanding research assistance by Ben Harack and Maximilian Negele and for very helpful feedback from Eric Gartzke, Nadiya Kostyuk, and attendees of our presentations at the Future of Humanity Institute, the Centre for the Governance of AI, Dartmouth College, and the American Political Science Association 2021 Annual Meeting.

To make a discovery is not necessarily the same as to understand a discovery. - Abraham Pais[1]

In the summer of 1942, a group of physicists held a series of secret meetings at the University of California, Berkeley, over the development of the nuclear bomb. One of those physicists, Edward Teller, realized that a reaction had the potential to ignite all the hydrogen in the oceans or nitrogen in the atmosphere, destroying all complex life on Earth. Though Teller's calculations were later understood to be flawed, the physicists remained worried about the potential for catastrophe up until the Trinity test was conducted in 1945.[2] And yet, they conducted the test. Part of the decision to implement the atomic bomb was due to the US's fear that Nazi Germany was developing their own bomb. What factors influence decisions to implement risky technologies in competitive contexts?

Technological arms races often create pressure on governments to take risks that affect not only their own citizens, but also their adversaries' citizens and even those of neutral countries. For example, criminal groups have stolen and used cyberweapons developed by the US government: by creating these tools, the US inadvertently created new dangers to its own citizens and to people around the world.[3] During the Cold War, the US and the USSR took risks that could have been catastrophic. Infectious pathogens escaped from Soviet biological weapons labs several times.[4] Humanity came perilously close to nuclear war on numerous occasions when an American or Soviet early-warning system falsely reported that the other side had launched a nuclear attack.[5]

Today, the United States, China, and other great powers are engaged in an increasingly dangerous and destabilizing race to develop new technologies useful for geopolitical competition, particularly applications of artificial intelligence (AI) to areas from robotic weapons to automated intelligence analysis.[6] For example, both the US and Chinese governments

[1] *Inward Bound: Of Matter and Forces in the Physical World*, 1988
[2] Ord 2020, 90-93.
[3] Allen and Chan 2017, 26
[4] Allen and Chan 2017, 110
[5] Borrie 2019
[6] State Council of China 2017; National Counterintelligence and Security Center 2021; Schmidt et al. 2021.

have implemented policies aimed to achieve a lead in access to the most advanced semiconductors, which are critical for developing AI applications. In 2022, the US passed the "CHIPS and Science Act," which subsidizes companies manufacturing semiconductors on US soil and bars recipients from investing in new factories to produce the most advanced semiconductors in China, Russia, or other "countries of concern." The US also tightened restrictions on the export of semiconductors to China, in an effort to prevent the Chinese military from using them to develop new weapons systems. The Chinese state has been investing heavily in domestic chip production and in 2022, a state-funded company started manufacturing semiconductors that matched the most cutting-edge technology in the world. What consequences will these competing policies have for global security?[7]

AI has vast potential to improve human welfare in areas such as medicine, education, and many others.[8] Even in the security realm, some emerging applications of AI could improve safety by, for instance, helping to evacuate wounded soldiers or monitor compliance with the laws of war.[9] However, there is growing concern among scholars that some AI-powered tools of statecraft could carry substantial global risks if they are deployed hastily, before sufficient research and testing has identified safe ways to design and use them.[10] Geist and Lohn (2018, 21) argue that "the riskiest periods will occur immediately after AI enables a new capability, such as tracking and targeting or decision support about escalation. During this break-in period, errors and misunderstandings are relatively likely. With time and increased technological progress, those risks would be expected to diminish." Intense security competition can incentivize states to deploy new technologies before passing through this perilous "break-in period."

There is a longstanding literature documenting that externalities and competition lead to greater risk-taking. Together, they can create a "race-to-the-bottom" dynamic, in which actors rush to deploy risky technologies, knowing their competitors may reap the benefits of

[7] Zhong and Li 2022; McKinnon and Fitch 2022; Sanger 2022; Swanson 2022
[8] Bommasani et al. 2021
[9] von Braun et al. 2021, 7.
[10] Allen and Chan 2017; Farquhar et al. 2017; von Braun et al. 2021

deploying first if they do not.[11] What has not been adequately studied are the conditions under which these dynamics are more and less severe. In particular, in the context of technology races between states, how do factors like the acuteness of safety-performance tradeoffs or actors' relative positions in the race lead them to take on more or less risk?

To answer these questions, we construct a model of a race between two states to deploy a technology that poses a risk to both. We examine how three factors influence their willingness to cut corners on safety in order to win the race: the relative technical knowledge of the competitors; the severity of the safety-performance tradeoff; and the degree of enmity between them.

In the model, two actors compete to develop and use a technology first. Players' technical knowledge increases over time probabilistically, so that the initially less knowledgeable player may catch up to or overtake the leader. Each period, the players choose whether to build and use the technology or wait until they understand it better. If they decide not to wait, they can implement the most capable version of the technology possible given their knowledge level, or a version that is less capable but which they understand better and is thus safer. If a player implements a more capable version of the technology, it is more likely to win the race, but also more likely to cause a disaster that harms both players. Thus, the players weigh a tradeoff between performance and safety in making implementation decisions. This tradeoff is a key feature of these strategic contexts.

We establish some counter-intuitive findings. First, we characterize how the safety-performance tradeoff affects the probability of an adverse outcome that harms all players. There are two competing effects. Conditional on the version of the technology a player implements, the direct effect of improving the safety of that version is of course to reduce

[11] See for example: Kahler 1998, Berger 2000, Murphy et al. 2004, Cai and Treisman 2005, Prakash and Potoski 2006. Similar strategic dynamics play out in other contexts, such as competition between technology firms. A recent example comes from Meta, the company formerly known as Facebook, Inc. In order to protect its declining market share, it knowingly employed algorithms on Facebook that promoted anger-provoking content, potentially furthering societal polarization (see: Timberg, Craig, "New whistleblower claims Facebook allowed hate, illegal activity to go unchecked" *Washington Post*, 22 October, 2021). The decision to do so was a risk not only to the company itself, but also to the industry's largely unregulated status and perhaps to the effectiveness of democracy.

the total risk. However, there is also a risk compensation effect: the safety improvement increases incentives to implement recklessly. This leads to a counter-intuitive result: sometimes investments in safety can increase risk, by inducing actors to deploy more dangerous versions of a technology than they would have otherwise.[12] Yet, past a certain level of safety of the technology, any further increases in safety decrease total risk. These monotonically positive returns to safety improvements set in at lower levels of safety when enmity is *higher*. This is because at low enmity, a larger increase in the safety of the technology at a given performance level is required to incentivize an actor to implement at that performance level, rather than at a lower performance level. Thus, safety insights decrease overall risk only when the technology is already sufficiently safe, and the level at which a technology is already sufficiently safe decreases in enmity. Thus, when enmity is higher, safety insights are more likely to reduce overall risk; when enmity is lower, safety insights must be more effective before they reduce, rather than increase, overall risk.

Second, a close race can be either safer or more dangerous than a race in which one player is behind, depending on the enmity between the players and the laggard's capability. In particular, if enmity is only moderate or the laggard is so far behind that it has little opportunity to win even if it cuts corners, the shared risk is lower than if the laggard catches up to the leader. However, when enmity is high enough and the laggard is sufficiently capable, the laggard is willing to cut corners, increasing the risk of a shared disaster. In this case, the risk of a bad outcome for both players *decreases* when the laggard catches up.

Thus, capable but frustrated laggards are dangerously motivated to take risks - and may induce technology leaders to take greater risks in turn. As an analogy, consider runners near the end of a race. Racers who are neck and neck, even if they are desperate to win, are unlikely to cheat by cutting corners if cheating stands a sufficient chance of being caught. Racers who are far behind are likely to accept that they cannot win and give up. But racers

[12]Note that this risk compensation effect is different from moral hazard. Risk compensation is caused by decreasing the probability that a given action will cause harm (in this case, the probability that implementing a particular version of the technology will cause a disaster), whereas moral hazard is created by shifting the harm from a decision-maker to another party (Reynolds 2015).

who are not quite so far behind know that they have a chance of winning, but only if they cut corners.

Most of the parameters have effects on risk whose signs are conditional on the other parameters; not so with enmity. Enmity is the reason to cut corners; when higher enmity has an effect it is always to increase the risk of disaster.[13] If the laggard does not view its opponent's victory as catastrophic, it is less likely to engage in risky development to catch up. If the United States is developing a new weapon, it is unlikely that the UK would trade off safety for speed in an attempt to catch up. On the other hand, an adversary such as Russia that strongly prefers to develop the weapon before the US might be willing to trade off more risk in order to win the race.

The model is related to, but distinct from, existing models in the technology and arms race literatures. In contrast to most arms race models, the model below is a dynamic, infinite horizon model with a state variable - players' knowledge levels - that can grow over time. In contrast to models in the technology and patent race literature, we study a situation in which actors take joint risks and face a choice to implement a technology immediately or to wait until they learn how to reduce its risks. Thus, one actor may have an incentive to implement a risky technology before another has a chance, not just to win the race, but to prevent the risky corner-cutting behavior of the other actor. No previous study analyzes a race model that combines all of these features.

The next section characterizes our model relative to the existing literature. We then present the model formally and discuss its implications, including how those interested in maximizing general welfare can intervene depending on race characteristics. We also characterize the ways in which the model generalizes to other cases through three extensions.

[13] Most theoretical paradigms in international relations conceive of enmity in a similar fashion to the way that we understand it here. See Goertz and Diehl 1995 and Hensel, Goertz and Diehl 2000 for definitions.

Racing Under Risk

Political leaders frequently face tradeoffs between increasing their state's power over its rivals and reducing global risks. Should they deploy a powerful new weapon as soon as possible or hold off until they have learned more about how to use the technology safely? If they deploy it, should they design and use it in the way that maximizes the likelihood of prevailing in a crisis, or in a way that makes makes a shared catastrophe less likely but a stalemate more likely?

For example, nuclear-armed adversaries face a shared risk of an accidental or unauthorized launch triggering a nuclear war. Building nuclear bombs involves accepting this risk in return for the ability to deter or compel rival nations.[14] Moreover, there are tradeoffs between military capability and mutual safety in *how* one deploys nuclear weapons. For instance, President Eisenhower delegated authority to American commanders in Europe to use nuclear weapons in the event of a Soviet invasion of NATO countries. This decision was meant to ensure they could stop an invasion quickly, particularly if a communications breakdown prevented them from communicating with the president. Yet Eisenhower was aware that delegating control over nuclear weapons raised the risk that "something foolish down the chain of command" could provoke an unintended nuclear war.[15] Similar capability-safety tradeoffs apply to decisions about whether to keep nuclear forces on high alert[16] and whether to use "permissive action links," mechanisms that prevent the launch or detonation of a nuclear warhead unless the operator enters a secret code they would only receive during a crisis.[17]

[14] Even if a country's adversaries do not possess nuclear weapons, acquiring nuclear weapons may tempt them to do so as well, making an accidental nuclear exchange possible.

[15] See: Schlosser 2014. It appears that Eisenhower delegated nuclear authority solely to improve America's chances of winning a war with the Soviet Union, not to strengthen its ability to deter one, since as Schlosser notes, Eisenhower chose to keep the decision secret.

[16] See Intriligator and Brito 1985; Kristensen and McKinzie 2012.

[17] Schlosser 2014. In some cases, optimizing the design and use of a weapon for maximum competitive advantage over adversaries might also minimize shared risks along *some* dimensions. Even in such cases, there may still be a tradeoff between competitive advantage and avoiding shared risk along other dimensions. In any case, our claim is simply that significant tradeoffs between strategic advantage and shared safety are sufficiently common to be worth studying.

Great power competition over emerging technologies poses new global risks in the current era. For example, states are competing for military and political advantage through various applications of AI, in areas such as cyberwarfare, logistics, autonomous drones, analysis of intelligence data, and spreading disinformation.[18] As a shorthand, we will refer to all applications of AI by states to increase their relative power as "AI weapons," but as these examples show, such applications extend far beyond narrowly-defined military technologies.

Across a number of disciplines, a growing community has begun to identify substantial global risks from some forms of advanced AI.[19] Efforts by a state to develop and use AI weapons could end up harming both itself and its adversaries, as well as other states. These types of unintended consequences could result from malfunctioning weapons, proliferation, or systemic change in states' incentives to initiate or escalate conflict.[20]

As an example of malfunctioning risk, an AI advisor system that analyzes intelligence data to assess the military intentions of another state might mistakenly conclude that an attack is imminent and recommend a preemptive strike, leading to a war that would otherwise not have occurred.[21] AI poses significant proliferation risks too. For example, other states, terrorists, or criminals can steal AI cyberweapons, making all states more vulnerable.[23] Finally, AI weapons could change strategic incentives in dangerous ways. For example, a state with automated systems that could locate another country's nuclear weapons or

[18] See: Ayoub and Payne 2016; Allen and Chan 2017; Thomas 2020; Schmidt et al. 2021.

[19] See for example: Farquhar et al. 2017. Some of the risks from AI and other emerging technologies may even be existential. Bostrom 2002 is one of the first contemporary works to advocate for unifying the study of existential risks. See Bostrom and Cirkovic 2011 and Ord 2020 for current reviews of natural and anthropogenic existential risks.

[20] This categorization of AI risks is based on the broader typology proposed by Zwetsloot and Dafoe 2019, who sort the unintended harms of AI into "accident," "misuse," and "structural" risks.

[21] See Geist and Lohn 2018 and Price, Walker and Wiley 2018. Price, Walker and Wiley (2018) note that accidental "flash wars" would be even more likely to break out if both sides have automated systems for strategic analysis, which could interact in unexpected ways, just as automated stock trading algorithms have produced selling sprees dubbed "flash crashes." While any technology can malfunction, the increasing complexity of AI systems makes it particularly hard to predict how they will act in novel situations.[22] This risk is compounded by the speed and scale that automation allows: humans may not have time to stop an AI before it inflicts substantial damage (Altmann and Sauer 2017; Danzig 2018; Price, Walker and Wiley 2018; Zwetsloot and Dafoe 2019; Rudner and Toner 2021).

[23] As Allen and Chan (2017) point out, stealing the design for a physical weapon does not immediately enable the thieves to build that weapon if they lack the necessary materials or manufacturing capability. But if they steal the computer code for a cyberweapon, they have stolen the weapon itself.

command-and-control infrastructure might be able to preemptively destroy an adversary's ability to retaliate against a nuclear attack. In a crisis, the opposing state might choose to launch its nuclear weapons while it still can.[24]

There may be significant first-mover advantages in deploying some kinds of AI. Because AI lends itself to rapidly scaling up, breakthroughs in AI weapons technology may lead to dramatic shifts in the balance of power.[25] These potential strategic advantages could create powerful incentives for states to field cutting-edge AI weapons before they fully understand their risks.

Previous analyses of technology and arms races offer useful analogies in addressing these questions, but these literatures do not capture essential features of risky technology races. Most strategic arms race models do not include a state variable in spite of the centrality of the growth in arms over time to the strategic context.[26] Fearon 2011 studies a dynamic, two-actor model with state variables representing accumulated arms, but arming is deterministic.[27]

The technological race model we analyze in this article also draws from the literature on patent and innovation races.[28] Our model shares the dynamic nature of these models and the uncertainty in the arrival time of innovations (Budd, Harris and Vickers 1993; Bimpikis, Ehsani and Mostagir 2019). Like existing innovation race models, there is a benefit to the actor who reaches a threshold first and a cost to the actors who do not. This might be thought of as a state achieving a long-term strategic advantage or a short-term advantage that is sufficiently important to drive a technological arms race. Like Bimpikis, Ehsani and Mostagir 2019, we describe how the research progress of opponents conditions players' incentives to take risks. Like some others, we include a state variable that measures players' level of progress in the race, finding that large gaps can discourage laggards (Scotchmer

[24] Moore Geist 2016; Horowitz 2019; Zwetsloot and Dafoe 2019.
[25] Horowitz 2018; Future of Life Institute 2021.
[26] Examples include: Axelrod 1984; Downs and Rocke 1990; Powell 1993; Jackson and Morelli 2008; Fearon 2018.
[27] Bas and Coe 2016 study a dynamic context in which one side attempts to master a technology while the other side attempts to deter mastery, but is not racing itself.
[28] See Langinier and Moschini 2002 for a review of this literature.

and Green 1990; Bimpikis, Ehsani and Mostagir 2019). Unlike this literature, however, we allow actors to choose between waiting for research that allows safer implementation of technologies and cutting corners to implement a risky technology now. We also differ from the patent race literature in our welfare analysis and in how we conceive of the opportunities for actors to mitigate joint risks. In much of the patent race literature, welfare loss comes from the fact that laggards invest in R&D without winning the prize, leading to deadweight loss (Denicolo 2000; Rockett 2010). In our model, by contrast, welfare loss to both players is driven by the fact that one or both players can cut corners by deploying technologies that pose shared risks.

A Model of Implementation Risk

The model is a stochastic game in which players compete for a decisive strategic (military or political) advantage that comes with being the first to successfully implement a particular technology. However, in attempting to implement this new technology, they run the risk of causing a disaster that imposes costs on both. The players face a tradeoff between winning the competition and avoiding a shared disaster. As the game goes on, players acquire new technical knowledge that allows them to compete more effectively or more safely.

In each period, players choose whether to implement the technology and if so, which version of the technology to implement. Implementation represents any effort by states to design and build a form of the technology and attempt to use it to gain a strategic advantage. For instance, a state might decide to design a new kind of fully autonomous swarming drone, mass produce them, and threaten to use them in order to compel or deter a rival. Or a state might seek to create a new cyberweapon and attack another power's financial institutions to reduce the rival's relative power.

Implementation attempts can be successful or unsuccessful. As an example of an unsuccessful implementation attempt, a state might attack its rival with a new weapon only

to discover that it is ineffective in combat because of a design flaw or because the target state has an effective defensive technology to counter it. Implementation may also fail for non-technical reasons. A military's strategy for integrating the weapon into its operations may be ineffective or political leaders might underestimate their opponents' resolve such that the new weapon fails to deter them. If players implement the technology, they choose among different versions of it: e.g. different technical designs or operational procedures for deploying the technology. The versions vary in terms of their **capability**, which we define as the likelihood that implementing that version will result in the implementer achieving the strategic advantage, i.e. winning the competition.[29]

The state of the game in a given period is the two players' levels of knowledge about the technology and a player's knowledge determines which versions of the technology it can implement. At the beginning of each period, each player observes its own knowledge level and and that of its opponent. In period $t \in \{1, 2, ..\}$ each player $i \in \{1, 2\}$ has knowledge level $K_{i,t} \in \{1, 2, 3\}$. The greater a player's knowledge, the greater the maximum capability of the versions of the technology they can implement. (For example, over time the US and the Soviet Union learned how to build increasingly powerful nuclear warheads.) At the end of each period, as long as they have not reached the highest knowledge level ($K_{i,t} = 3$), each player has an independent probability $\alpha \in (0, 1)$ of advancing to the next level of knowledge.

After observing their knowledge levels, the players simultaneously choose which versions of the technology to implement, if any. Specifically, each player chooses a capability level $C_{i,t} \in \{0, ..., K_{i,t}\}$, where higher values of $C_{i,t}$ correspond to more capable versions of the technology and $C_{i,t} = 0$ represents not implementing any version.

There are three mutually-exclusive possible outcomes in each period. The first possibility is that one of the players succeeds in implementing the technology and wins the competition for the strategic advantage, ending the game. The second possibility is that one of the players

[29] The exact probability that a player achieves the strategic advantage depends on the other player's implementation choice in that period, as we describe below. However, holing the opponent's action fixed, implementing a more capable version of the technology always increases a player's chance of winning in that period.

causes a disaster, which also ends the game. Disasters represent unintended outcomes that are catastrophic for both rival powers, such as an unintended nuclear war or a weapon of mass destruction falling into the hands of a terrorist group. The final possibility is that neither player successfully implements or causes a disaster and the game continues to the next period. We call this possibility a **status quo outcome** and it can occur either because neither play tried to implement the technology or because implementation was unsuccessful.

If a player reaches the highest knowledge level ($K_{i,t} = 3$), it has fully mastered the technology in the sense that it cannot cause a disaster or experience a failed implementation. We discuss this case in further detail below. Otherwise, if a player is still at knowledge level $K_{i,t} = 1$ or 2, it faces a safety-performance tradeoff. Implementing a more capable version of the technology decreases the player's chance of losing the competition for strategic advantage but increases the chances that the player will cause a disaster.

Specifically, the probability that a player causes a disaster is decreasing in the gap between its knowledge level and the capability level at which it implements. This assumption reflects both the safety-performance tradeoff at any particular point in time and the tendency over time for states to gain insights into how to manage the risks of the technology. For a given level of capability, states learn how to deploy increasingly safer versions of the technology. Conversely, for any given level of risk, states will be able to deploy increasingly more capable versions. The most capable version of a technology available to a state at a given time will often be the most cutting-edge one; for these versions, states have had less time to conduct tests, assess risks, and devise mechanisms and operating procedures to improve safety.

Formally, if $K_{i,t} = K < 3$ and $C_{i,t} = C > 0$, the probability of the implementation causing a disaster is given by a function δ_{K-C}, where $0 < \delta_1 < \delta_0 < 1$. We refer to δ_0 and δ_1 as the **full implementation risk** and the **partial implementation risk**, respectively. The probability that a player does not implement successfully but does not cause a disaster is decreasing in the level of implementation, and is given by the function σ_C, where $0 < \sigma_2 <$

$\sigma_1 < \sigma_0 = 1$.[30]

The outcome at the end of each period also depends on the relative capability levels at which players try to implement. If Player i implements at a higher capability level than Player j ($C_{i,t} > C_{j,t}$), the higher-implementing player has the first shot to win or cause a disaster. If that player implements unsuccessfully but does not cause a disaster, the lower-implementing player then has a chance to win or cause a disaster.[31] If both players implement at the same level, then each has a 50% chance of getting the "first shot" to win or cause a disaster.[32]

Once a player has reached the highest knowledge level, it cannot cause a disaster or experience a failed implementation. Conditional on getting a "shot" at winning, if Player i has knowledge $K_{i,t} = 3$ and chooses to implement at a non-zero capability level ($C_{i,t} \geq 1$), then it wins the competition with probability 1.[33]

We make some additional assumptions to focus on interesting cases. First, we restrict the parameter values in ways that ensure that a player's chance of winning in any given period is weakly increasing in the capability level at which they try to implement. Namely, we assume that $\sigma_1 + \delta_0 < 1$ and $\sigma_1 - \sigma_2 > \delta_0 - \delta_1$. Furthermore, when neither player is at

[30] Although we model how knowledge about the technology affects capabilities and risk, similar dynamics likely apply to many other factors that influence the combinations of performance and safety a state can achieve when deploying a technology. For instance, a state's abilities in AI domains depend on their access not only to cutting-edge algorithms, but also to well-trained and experienced engineers, powerful computer hardware, and large training datasets (Ding 2018). Economic, political, and other social factors will also influence the extent to which AI capabilities translate into strategic advantage.

[31] As noted, the players move simultaneously. The higher-implementing player gets the "first shot" at ending the game only in the sense that, by winning the competition, the player who deploys a more powerful technology preempts the possibility of a disaster by the other player. Relaxing this assumption, so that implementing at a higher capability does not reduce the chance that the other player causes a disaster in that period, should not qualitatively change our results. Winning would still eliminate the possibility of the other player causing a disaster in a *future* period, so a similar incentive to preempt a disaster by the other player by implementing at a high capability would remain.

[32] We define a variable D_t, such that $D_t = 1$ if either player causes a disaster in period t and $D_t = 0$ if neither player causes a disaster in period t.

[33] As in the cases in which $K_{i,t} < 3$, Player i gets the first shot at winning with probability 1 if it implements at a higher level than Player j ($C_{i,t} > C_{j,t}$) and with probability 0.5 if they implement at the same non-zero level. If Player j gets the first shot, Player i will win if and only if Player j implements unsuccessfully (but doesn't cause a disaster) and Player i has chosen a non-zero implementation level. If Player i does not try to implement ($C_{i,t} = 0$), it cannot win in that turn, and the game continues if Player j also does not implement or implements unsuccessfully.

the highest knowledge level, these restrictions imply that the chance of winning is *strictly* increasing in the capability level at which a player implements.[34]

The payoffs are as follows. At the end of the game, if a disaster occurs, each player gets a payoff of -1. If there is no disaster, the winning player gets a payoff of 0 and the losing player gets a payoff of $-e$, where $0 < e < 1$. We refer to e as the enmity between the players. The assumption that $e < 1$ is an important scope condition for our analysis: we are focused on situations in which the shared disaster is worse for each player than losing the competition for strategic advantage. For example, the leaders of two rival states might both prefer a world in which the other is a global hegemon to a world in which both countries have been annihilated in a nuclear war. If neither player ever causes a disaster or successfully implements (such that the game continues forever), each gets a payoff of $-q$, where $0 < q < e$.[35]

Results

When is a technology race of this sort most dangerous? What factors determine the likelihood that a race will eventually end with one of the players causing a disaster, which we call the **overall disaster risk**? In this section we present two counter-intuitive insights that emerge from the model about what makes a race dangerous. (1) Improvements in the safety of a technology can make a race riskier overall. (2) A race can be safer when it is neck-and-neck, compared to when one player is more technologically advanced than the other. In each of the following two subsections, we discuss one of these findings and the factors that determine when the finding does and does not hold.

In the first subsection, we show that marginal safety improvements can sometimes be self-defeating because of risk compensation. Reducing the probability that an actor will cause a disaster if they implement the most powerful technology they are capable of can induce

[34] If Player i is at the highest knowledge level, its chance of winning is 1 if it implements at any capability level such that $C_{i,t} > C_{j,t}$.

[35] The game can only continue forever if neither player ever tries to implement after reaching the highest knowledge level.

them to implement it when they would otherwise have implemented a less capable but safer version. Moreover, one player's decision to implement at a higher level creates pressure on the other to take greater risks as well. Thus, improvements to safety can actually increase the overall probability that a disaster eventually occurs.

The second subsection explores when a close race is safer or more dangerous than one with a clear leader. We find that, in a wide array of circumstances, a racer will behave more recklessly when they are less technologically advanced than their rival, and exercise greater caution if they catch up. Thus, when there is a leader and a laggard, the overall disaster risk can be higher than when the race is neck-and-neck. In particular, a close race tends to be safer than a race with a leader and a laggard when enmity between the players is high and when even the less-advanced technology available to the laggard is relatively effective.

We analyze the set of symmetric pure-strategy Markov-perfect equilibria. By focusing on Markov-perfect equilibria, we can investigate interventions that involve altering material factors: the costs and benefits competitors face in different outcomes and the likelihood of the outcomes conditional on the competitors' choices.[36]

In *all* Nash equilibria, as soon as one or both players reach the highest knowledge level ($K_{i,t} = 3$), the game will end in victory for one of them with certainty.[37] So players' expected payoffs and best responses in all states of the game in which neither has reached the highest knowledge level ($\max\{K_{1,t}, K_{2,t}\} < 3$) cannot depend on which of the subgame equilibria they play in any subgame in which one or both players *have* reached the highest knowledge level ($\max\{K_{1,t}, K_{2,t}\} = 3$). We therefore assume, without affecting any of our results, that

[36] The restriction to Markovian strategies is fairly standard in the arms race literature. See for example: Jackson and Morelli (2008) and Fearon (2011). Considering asymmetric, mixed-strategy, and non-Markov equilibria in technology race models would be a valuable direction for future research.

[37] When exactly one player has knowledge $K_{i,t} = 3$, there is no Nash equilibrium in which the leading player implements at a capability level less than or equal to that played by the lagging player. Doing so would yield a payoff of $-e$, $-0.5e$, or $-q$, whereas the leader can guarantee victory and a payoff of 0 by choosing $C_{i,t} = 3$. If *both* players are at the highest knowledge level ($K_{1,t} = K_{2,t} = 3$), the *only* Nash equilibrium is for both to implement fully ($C_{1,t} = C_{2,t} = 3$). In any strategy profile in which both players implemented below capability level 3, either player could profitably deviate to $C_{i,t} = 3$, guaranteeing a win and the highest payoff, rather than a payoff $-e$, $-0.5e$, or $-q$. In any strategy profile in which one player implemented at $C_{j,t} = 3$ and the other player implemented $C_{i,t} < 3$, the latter could profitably deviate to $C_{i,t} = 3$ and have an expected payoff of $-0.5e$ rather than $-e$.

both players always implement fully ($C_{i,t} = K_{i,t}$) once either of them reaches the highest knowledge level ($\max\{K_{1,t}, K_{2,t}\} = 3$).[38]

We denote a player's equilibrium implementation level in any given period as a function of their position in the race: $C^{\bigstar}(x, y)$. When a player (i) has knowledge $K_{i,t} = x \in \{1, 2, 3\}$, and the other player ($j$) has knowledge $K_{j,t} = y \in \{1, 2, 3\}$, player i chooses implementation level $C^{\bigstar}(x, y) \in \{0, ..., x\}$. Since we assume that $C^{\bigstar}(3, y) = 3 \; \forall \; y \in \{1, 2, 3\}$ and $C^{\bigstar}(x, 3) = x \; \forall \; x \in \{1, 2\}$, we will differentiate between strategies simply in terms of what a player does in states of the race when neither player has reached knowledge level 3. We denote strategies as: $C^{\bigstar} = (a, b, c)$, representing the implementation levels a player chooses when it is leading, lagging, or tied, respectively. $a \in \{0, 1, 2\}$ is the level of implementation players choose when they are ahead with knowledge level 2 while the other player is at knowledge level 1: $C^{\bigstar}(2, 1) = a$. $b \in \{0, 1\}$ is the implementation level they choose when they are behind at knowledge level 1 while the other player is at knowledge level 2: $C^{\bigstar}(1, 2) = b$. $c \in \{0, 1, 2\}$ is the implementation level they choose if the initial laggard has caught up to the leader at knowledge level 2: $C^{\bigstar}(2, 2) = c$. Because we restrict our analysis to symmetric equilibria, we will also use this notation to represent equilibria. We denote the overall disaster risk as $P^{D}_{x,y}(C^{\bigstar})$, because it is conditional on the equilibrium and on the current state of the race.[39]

The model sometimes has multiple symmetric pure-strategy Markov-perfect equilibria. We focus our analysis on the equilibrium with the lowest probability that a disaster occurs before the race ends, given players' initial knowledge levels. We refer to this probability as the **minimum overall disaster risk**, denoted $\widehat{P}^{D}_{x,y}$.[40]

[38] For all parameter values consistent with our assumptions, this subgame-perfect equilibrium exists in all subgames that begin with one or both players at the highest knowledge level. For a who player starts at knowledge level 3, it is a weakly dominant strategy for that player to always choose $C_{i,t} = 3$. If *only* one player is at knowledge level 3 ($K_{i,t} = 3$, $K_{j,t} \in \{1, 2\}$) and the leader implements at capability level 3 ($C_{i,t} = 3$), the lagging player gets a payoff of $-e$ no matter what action it chooses, so it does not have a profitable deviation from full implementation.

[39] The overall disaster risk is not to be confused with the probability of a disaster occurring on a *single* implementation attempt at a level C by a player with a knowledge level K. The latter probability is simply δ_{K-C}, whereas $P^{D}_{x,y}(C^{\bigstar}) = Prob\{\exists \; T \geq t : D_T = 1 | C^{\bigstar}, K_{i,t} = x, K_{j,t} = y\}$, where T is the period in which the game ends. (Note that T is a random variable in $\{t, t+1, ...\}$.)

[40] Let M denote the set of symmetric pure-strategy Markov-perfect equilibria: $\widehat{P}^{D}_{x,y} = \min_{C^{\bigstar} \in M} P^{D}_{x,y}(C^{\bigstar})$.

Of course there may be some empirical technological arms races in which actors fail to coordinate on feasible lower-risk strategies.[41] Nonetheless, by focusing on the minimum-risk equilibria, we are able to understand the counter-intuitive dynamics of risk in best-case scenarios.

Improving safety can increase risk.

It might seem that improving the safety of a technology is always a good thing. But the model shows that in fact safety improvements can sometimes be counterproductive. Making the most powerful version of a technology safer can incentivize rivals in a race to use it, when they would otherwise exercise caution by deploying a less capable version that they understood better. On balance, improving the safety of racers' most powerful technologies can therefore increase the overall disaster risk.

Figures 1 and 2 illustrate this point. We focus in these charts on the case where the players start in what we call the "gap state," when one player leads with knowledge $K_{i,1} = 2$ and the other player lags behind at $K_{j,1} = 1$. In each chart, the horizontal axis is the probability of causing a disaster (δ_0) each time a player implements "fully", deploying the most powerful version of the technology it can, given its current knowledge.[42] Recall that in the model the risk of a disaster is solely a function of the gap between a player's knowledge and the level at which it implements, as long as it has not reached the highest knowledge level. So reducing the full implementation risk means reducing the risk of implementing at level 2 ($C_{i,t} = 2$) for a player with knowledge 2 ($K_{i,t} = 2$) and reducing the risk of implementing at level 1 ($C_{i,t} = 1$)for a player with knowledge 1 ($K_{i,t} = 1$). The vertical axis in each chart is the minimum overall disaster risk, conditional on the race being in the gap state ($\widehat{P}^D_{2,1}$). In

[41] In numerical computations, we find that the largest differences between the minimum and maximum disaster risks occur when the incentives to trade off safety for performance are in a middle range: in these cases players are willing to refrain from riskier levels of implementation, if and only if they expect the other player to do so.

[42] Full implementation risk is bounded below at δ_1, the probability of causing a disaster when a player implements one level below their maximum capability level, which we hold fixed at 0.05 in both charts. It is bounded from above by the assumption that $\sigma_1 - \sigma_2 > \delta_0 - \delta_1$. We set $\sigma_1 = 0.5$ and $\sigma_2 = 0.05$, so δ_0 must be less than 0.5. (This range also fulfills the assumption that $\delta_0 + \sigma_1 < 1$.) We also hold fixed $\alpha = 0.25$.

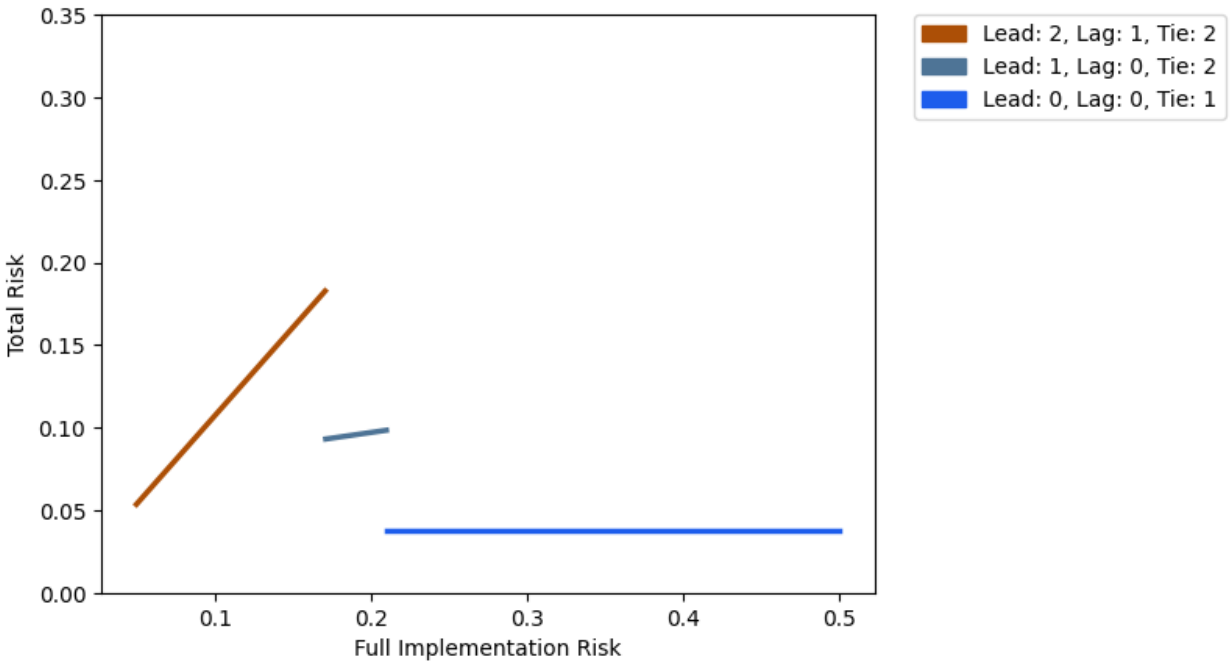

Figure 1: Minimum overall disaster risk at gap state versus full implementation risk (low enmity)

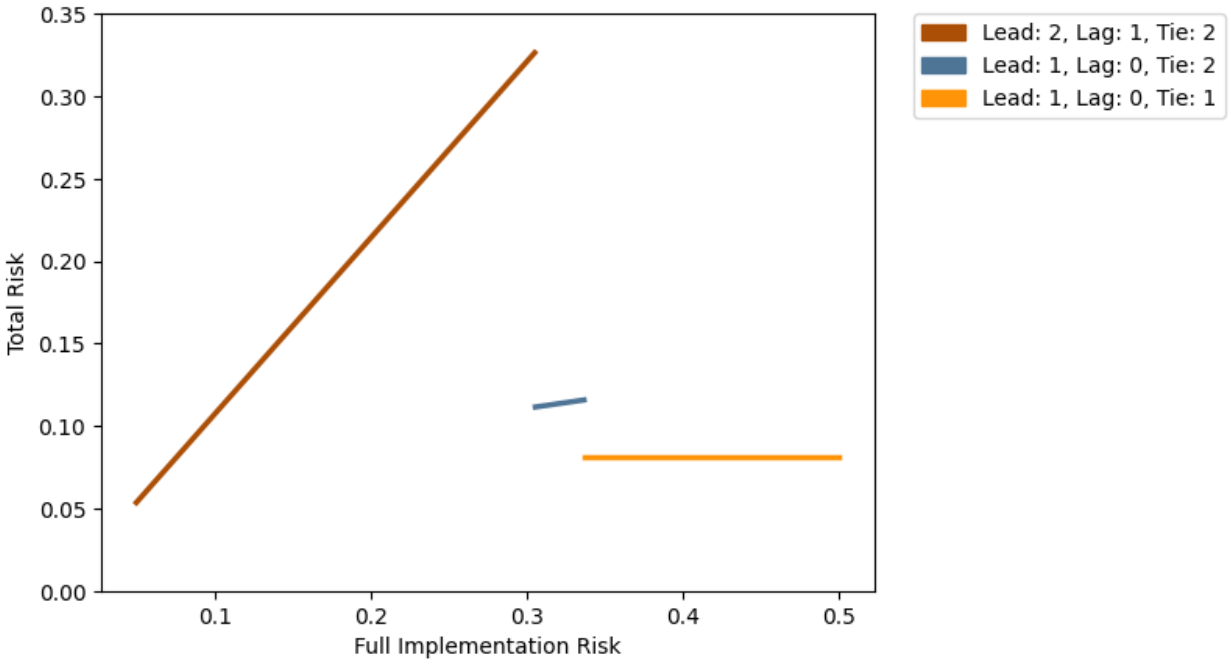

Figure 2: Minimum overall disaster risk at gap state versus full implementation risk (high enmity)

Figure 1, we hold enmity (e) fixed at 0.3, while in Figure 2 we hold enmity at 0.6. Each line segment corresponds to a different equilibrium: the one with the lowest overall disaster risk among all equilibria that exist at a given level of full implementation risk. The equilibria are described in terms of the implementation levels chosen when a player is in the lead (at knowledge level 2), lagging behind (at knowledge level 1), or tied (at knowledge level 2).

The minimum overall disaster risk is lowest when full implementation risk is either very low or very high. This non-monotonicity is driven by the two countervailing effects of full implementation risk on overall disaster risk. The first-order effect of increasing full implementation risk is to increase overall risk: holding the players' strategies constant, if players implement fully at some point in the race, the race is more likely to end in disaster as the probability of causing a disaster each time they implement fully goes up. The second-order effect of increasing full implementation risk, however, is to reduce the overall disaster risk: if players just slightly prefer to implement fully at some point in the race, making full implementation a little bit riskier can induce them to switch to implementing partially or not implementing at all. The first-order effect thus occurs within a given equilibrium. The second-order effect occurs across equilibria: when full implementation risk increases, a strategy profile with lower implementation levels that was not previously an equilibrium now is an equilibrium and becomes the new equilibrium with minimal overall risk.

The slope of each line reflects the first-order effect. The leftmost line of each graph shows the overall disaster risk when players implement fully whether they are in the lead, lagging behind, or tied. In this equilibrium, increasing full implementation risk raises overall disaster risk because the probability of a disaster every time any player implements is higher. The middle line in each graph shows the overall disaster risk when players only implement fully if they are tied. In this situation, overall disaster risk again increases in full implementation risk, because increases in the latter raise the chance that the race will end in a disaster if the laggard catches up. But the slope of this line is less steep, because full implementation risk does not affect the chance of the players causing a disaster while one player is still behind.

The rightmost line in each chart is flat: when full implementation risk is high enough, players do not implement fully at any point in the race, so further increases in full implementation risk have no effect on the overall disaster risk.

The discontinuous shifts from one line segment to another show the second-order effect at work. Consider the first chart, where enmity is relatively low. In this example, if full implementation risk is less than about 17%, the only equilibrium that exists is the one in which players always implement fully. However, when full implementation risk is between about 17% and 20%, an equilibrium exists in which the laggard does not implement because doing so would be too risky. Thus the minimum overall risk falls as full implementation risk increases past 17%. When full implementation is higher than about 20%, an even safer equilibrium exists, in which the leader also does not implement and players implement partially if they reach the what we call the "tie state," in which the initial laggard has caught up to the initial leader at knowledge level 2 ($K_{1,t} = K_{2,t} = 2$).[43] The dynamics are similar in Figure 2, in which enmity is higher.[44]

This effect also occurs when the players are tied. We summarize this non-monotonic relationship between full implementation risk and the minimum overall failure risk conditional on reaching the tie state in Proposition 1:

Proposition 1 *If low-tech performance (σ_1) is in an intermediate range and partial implementation risk (δ_1) is low enough, then there is some threshold $\delta_0^\bigstar$ and constant $\pi \in (0,1)$ such that:*[45]

[43] It may seem odd that the leader switches from partial implementation to non-implementation when full implementation risk increases, since the partial implementation risk has not changed. The shift in the leader's behavior in the gap state is driven by the shift in its expectation of what will happen if the laggard catches up. By choosing not to implement in the gap state, the leader raises the odds that it will reach the highest knowledge level before a disaster occurs, but it also raises the odds that the laggard will catch up before the race ends. If the leader anticipates that the players will implement only partially if they reach the tie state, the prospect of the laggard catching up is less costly for the leader in expectation, so the leader is willing to hold off on implementing.

[44] The only difference is that the leader continues to implement partially even as full implementation risk approaches its maximum, despite anticipating that the players will implement partially rather than fully if they reach the tie state.

[45] More formally, there exist thresholds $0 < \sigma_1^\bigstar < \sigma_1^{\bigstar\bigstar} < 1$ and $0 < \delta_1^\bigstar < \delta_0^\bigstar < \overline{\delta_0}$ such that $\delta_1 < \delta_1^\bigstar$ and $\sigma_1^\bigstar < \sigma_1 < \sigma_1^{\bigstar\bigstar}$ are sufficient conditions for the following conclusions.

$$\forall\ \delta_0 \in (\delta_1, \delta_0^{\bigstar}) : \widehat{P}_{2,2}^{D} \textit{ is strictly increasing in } \delta_0 \tag{1}$$

$$\forall\ \delta_0 \in [\delta_0^{\bigstar}, \overline{\delta_0}) : \widehat{P}_{2,2}^{D} = \pi < \lim_{\delta_0 \uparrow \delta_0^{\bigstar}} \widehat{P}_{2,2}^{D} \tag{2}$$

We prove Proposition 1 in Appendix A of the online supplemental materials. Numerical computation over a wide range of parameters indicates the result also holds conditional on the race being in the gap state (i.e. for $\widehat{P}_{2,1}^{D}$).

Intuition might suggest that policy interventions to improve the safety of a technology are most needed when it is very risky for competitors to implement their most powerful technology. However, in this model, reductions in full implementation risk only unambiguously reduce overall risk when full implementation risk is already so low that the players will implement fully at every stage in the race.

Comparing Figures 1 and 2 reveals another finding from this model that is quite robust: reducing enmity is always a good idea in the sense that minimum overall disaster risk is weakly increasing in enmity. At higher levels of enmity, the peak overall risk rises and shifts to the right. Since losing is costlier, full implementation risk has to be even higher for players to be willing to implement partially or not at all. This has the implication that, the lower enmity is, the less likely it is that reducing full implementation risk will reduce overall risk.

A close race is safer than one with a frustrated but capable laggard.

Intuition might suggest that rivals are most willing to cut corners when they are neck-and-neck. In fact, however, desperate laggards can be quite dangerous: if they try to implement, their limited knowledge makes doing so inherently risky. If they catch up, they might be more cautious, because they can achieve the same level of performance more safely. The overall disaster risk can therefore be higher when one player is behind, if (1) the cost of losing is high and (2) the laggard has a decent shot at winning by implementing its less-advanced

technology immediately.

Thus, under certain conditions that we detail below, the minimum overall disaster risk in the tie state is lower than in the gap state if and only if enmity and the probability of successful implementation with level 1 technology are both high enough. We refer to the latter as "low-tech performance,"[46] defined as $1 - \sigma_1$.[47] In equilibrium, laggards implement. If players are tied, they also implement the less-capable technology, but doing so carries a reduced risk (δ_1 rather than δ_0). Thus, the overall risk falls if the laggard catches up to the leader's knowledge level.[48]

Conversely, if enmity or low-tech performance is low, the laggard prefers not to implement unless and until it catches up to the leader, at which point both players implement partially. The leader in the gap state implements partially or not at all. Either way, the overall disaster risk rises if the laggard catches up.

Figure 3 illustrates this finding. On the horizontal axis is the level of low-tech performance and on the vertical axis is the level of enmity.[49] At the lowest levels of enmity there exists an equilibrium in which players do not implement until they reach the highest knowledge level, so both the tie risk and the gap risk are zero. For values of enmity in an intermediate range or for high values of enmity but low values of low-tech performance there are equilibria in which the laggard does not implement and the gap risk is lower than the tie risk.[50] At the highest levels of enmity and low-tech performance the laggard implements in all equilibria, and the tie risk is lower than the gap risk.[51]

[46] We do not mean to imply that a player with knowledge $K_{i,t} = 2$ who chooses $C_{i,t} = 1$ implements the *same* version of the technology as a player with knowledge $K_{i,t} = 1$. Rather, a player in the first scenario should be thought of as implementing a more sophisticated technology that achieves the same capability as a laggard's most powerful technology, but with a greater level of safety.

[47] $1 - \sigma_1$ is the probability that a player either wins *or* causes a disaster when it implements at level 1. However, in this section we consider the effect of σ_1 while holding δ_0 and δ_1 *fixed*, so variation in σ_1 only reflects variation in capability.

[48] Our simplifying assumption that knowledge levels are discrete limits us to comparing a gap to a tie. We are not suggesting that an exact tie between competitors in a race would be required to induce restraint, simply that they may become more cautious as the gap shrinks.

[49] The other parameters are set to: $\delta_0 = 0.11$, $\delta_1 = 0.01$, $\sigma_2 = 0.75$, and $\alpha = 0.01$.

[50] In all equilibria in this region, both players implement partially in the tie state, while the leading player in the gap state implements partially or not at all.

[51] In these equilibria, players implement partially when tied, while a leading player implements partially or

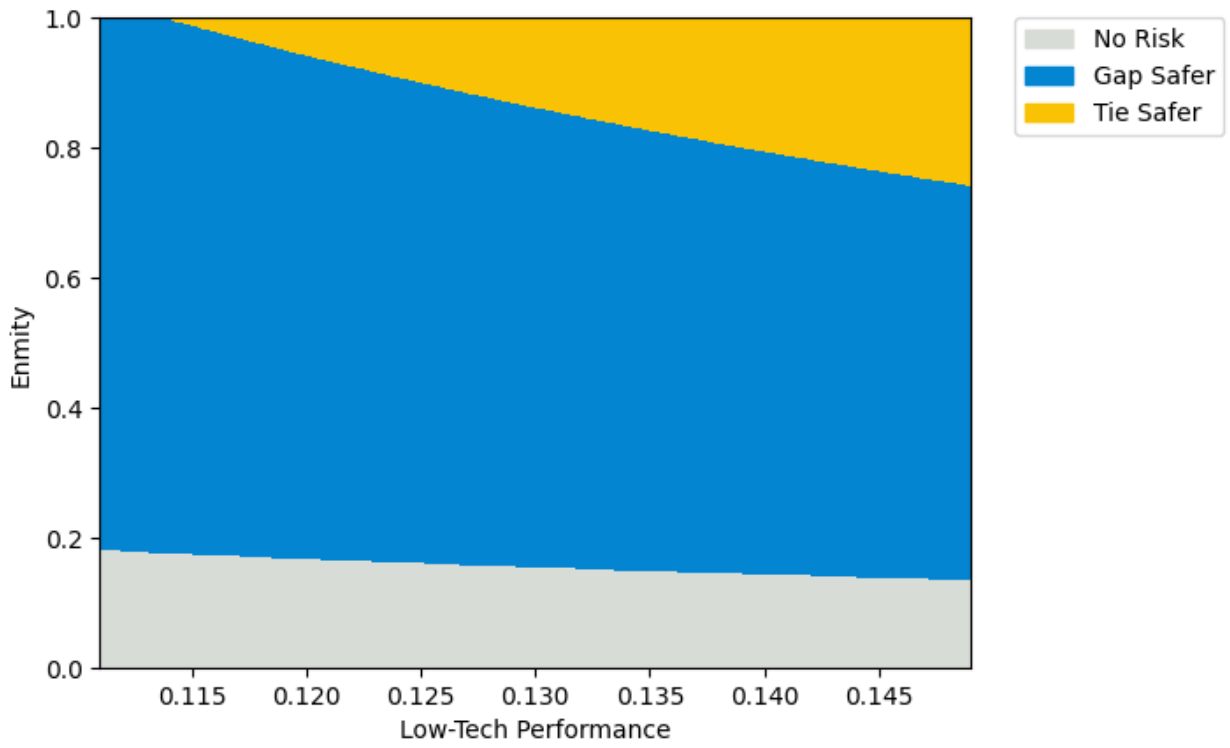

Figure 3: State of race with minimum overall disaster risk: gap (2,1) or tie (2,2)

Enmity and low-tech performance are substitutable in the laggard's expected payoffs: the higher the level of enmity, the lower the minimum level of low-tech performance at which the laggard prefers to implement and vice versa. This substitutability is reflected in the downward-sloping curve between the middle and upper regions. However, both enmity and low-tech performance must be above certain thresholds for the laggard to prefer to implement. If enmity is too low, the laggard will not implement no matter how high low-tech performance is, and if low-tech performance is too low, the laggard will not implement no matter how high enmity is.

For these equilibria to exist in their respective regions of the enmity/low-tech performance space, full implementation risk must be low enough that the laggard will implement for some sufficiently high enmity and low-tech performance. At the same time, full implementation risk must be *high* enough, and partial implementation risk low enough, that players will

fully.

implement partially (rather than fully) when tied, even at the highest level of enmity and the lowest level of low-tech performance. Based on this reasoning, Proposition 2 lays out sufficient conditions for the pattern that a tie is safer than a gap if and only if enmity and low-tech performance are high:

Proposition 2 *For low enough partial implementation risk (δ_1), intermediate full implementation risk (δ_0), and a rate of technological progress (α) that is either sufficiently low or sufficiently high, the minimum overall disaster risk in the tie state is less than the minimum overall disaster risk in the gap state if and only if enmity (e) and low-tech performance ($1-\sigma_1$) are sufficiently high. Specifically, for some threshold $\tau \in (2\delta_1, 1-\sigma_2-\delta_0+\delta_1)$:*

1. *If $e(1-\sigma_1) > \tau$, then $\widehat{P}^D_{2,1} > \widehat{P}^D_{2,2} > 0$ (gap riskier).*
2. *If $e(1-\sigma_1) \in (2\delta_1, \tau]$, then $\widehat{P}^D_{2,2} > \widehat{P}^D_{2,1} > 0$ (tie riskier).*
3. *If $e(1-\sigma_1) \leq 2\delta_1$, then $\widehat{P}^D_{2,1} = \widehat{P}^D_{2,2} = 0$ (no risk).*[52]

The conditions in Proposition 2 are sufficient but not necessary for the more general result that the tie state is safer than the gap state when enmity and low-tech performance are near their maximum values. In section 2 of supplemental Appendix C, we prove Proposition 3, which demonstrates that this finding also holds when partial implementation risk is *high* enough. In such cases, players do not implement at all in the tie state, so either the tie risk is lower than the gap risk — when enmity and low-tech performance are high — or the overall disaster risk is zero in both states.[53]

For a wide range of other parameter values, numerical calculations indicate that the tie risk is weakly lower than the gap risk for high enough enmity and low-tech performance. In particular, this result appears to hold as long as the difference between full and partial implementation risk is not too high and high-tech performance is not too low.[54]

[52] See the proof of Proposition 2 in supplemental Appendix B. We also found that enmity and low-tech performance must *both* exceed certain thresholds. We present and prove this corollary in section1 of supplemental Appendix C.

[53] When partial and full implementation risk are *both* high enough, players do not implement in *either* state, no matter how high enmity or low-tech performance.

[54] There is an interesting special case in which the result holds for a very different reason than under the

Extensions

The model above provides baseline results for the analysis of technology races with joint risks. In this section, we consider a variety of extensions.

Private Risks

We have described joint risks that states face in a variety of technology races. However, in many cases, when things go wrong, it is primarily the state that deployed the technology that is harmed.[55] The strategic dynamics that result are similar to the case of joint risks with one important difference: preempting other players from taking risky actions is not a reason for players to take riskier actions themselves. In spite of this difference, the key features of the strategic context remain: a safety-performance trade-off and a reward for whoever successfully deploys, which can motivate the racers to take greater risks than either would in the absence of competition.[56] For these reasons, we again find that Propositions 1 and 2 hold.[57]

Non-Existence of a Safe Knowledge Level

We assume in the baseline model that the third knowledge level is perfectly effective and safe. This assumption strengthens players' incentive to hold off on implementing until they reach the highest knowledge level. In this extension, we treat the third knowledge level in the same way that we do the others. That is, implementing at the limit of one's capability

conditions in Proposition 2. When full and partial implementation risk are both low, the tie is safer than the gap if and only if enmity and low-tech performance are high enough, even though players implement fully at all points in the race. In section 3 of Appendix C, we formalize and prove this result as Proposition 4. If the laggard catches up, the probability that someone reaches the (risk-free) highest knowledge level or ends the race by winning before a disaster occurs rises.

[55] For example, as the US military shifted rapidly from biplanes to jets during World War II, aviation accidents on American soil resulted in over 15,000 deaths (Danzig 2018).

[56] Even when there is no negative externality from a disaster, the externality the winner imposes on the loser (as captured by enmity) can drive a race to the bottom in which players take risks that leave them both worse off than if they could commit to implementing more cautiously.

[57] Appendix D of the supplemental materials provides a formal definition of the private risk version of the model and numerical examples of the results.

when a player is at the highest knowledge level is no safer than implementing fully at other knowledge levels, though it is more likely to result in a successful implementation of the technology.[58] Overall, we find that the results are largely similar. When one or both players is at knowledge level 3, we continue to find that overall disaster risk first rises and then falls in full implementation risk. Under conditions similar to those discussed for the baseline model, a race with a large gap is more dangerous than both a small gap and a tie when enmity and low-tech performance are high, and a large gap is safest when enmity and/or low-tech performance is low.

Larger Numbers of Racers

Propositions 1 and 2 also hold when this model is extended to more than two players. Moreover, versions of the model with three or more players allow us to compare different types of gaps with varying numbers of leaders. Paralleling the results from the two-player model, the gap state with one leader and two laggards (2,1,1) is safest when enmity or low-tech performance are relatively low, while a tie (2,2,2) is safest when enmity and low-tech performance are both high. For intermediate values of enmity and low-tech performance, the gap state with two leaders and one laggard (2,2,1) can be safer than either a tie or a gap state with just one leader.[59]

Discussion

Overall, the extensions provide reason to expect that the results of the baseline model are fairly robust. In particular, sometimes the shared risks of a technology race are higher when the competitors are closer together and sometimes they are higher when one competitor has a significant technological lead. One key factor that influences this is the level of enmity between the racers. When it is high, laggards cut corners on safety when they have the

[58] This extension is analyzed in detail in supplemental Appendix E.

[59] We illustrate these findings with numerical examples from the three-player model in Appendix F of the supplemental materials.

capability to do so. That is, they attempt to implement at the limit of their capabilities when they have little other chance of winning the race because they are far behind; if they were not so far behind, they would employ a less risky strategy. Thus, when distrust between great powers becomes more severe, we should expect laggards to run more substantial risks. When risks are joint, leaders will sometimes employ a riskier strategy as well in order to end the race before a disaster occurs.[60]

While we highlight these risks of unequal competition, it is also worth recalling that it is competition that produces risk in the first place. That means that when a laggard is so far behind that they have little chance of successfully deploying a particular technology even if they cut corners on safety, the overall risk is often at its lowest level. As a frustrated but *incapable* laggard starts to catch up, it will begin to act *more* recklessly.

A crucial set of assumptions in the model concern how the race ends. In any period, both a laggard and a leader have some chance to cross the finish line of successfully implementing the new technology. Importantly, the probability that one actor ends the race is not impacted by the relative position of the other actor. Thus, the race can end decisively when the players are neck and neck; a player can even when it is behind its rival technologically. We might also consider races that end endogenously when one side is *far enough* ahead of the other - in other words, when the laggard gives up. Such a race would be quite different from the one analyzed here, and likely more similar to the perpetual race analyzed in Hörner 2004. We expect in such a race that effort and corner cutting would be highest around this transition from a competitive to a non-competitive race because the marginal value of those actions is likely to be highest then. This is a useful avenue for future work.

This discussion highlights the role that endogenizing effort, and with it the speed of research progress, might play. Our model includes a fixed rate of progress and no costs of effort in order to focus on implementation decisions. Including effort in the model would

[60]The incentives for both players to implement sooner rather than later in the case of a "frustrated laggard" parallel the incentives for preventive attacks in deterrence theory and the spiral model, which reflect commitment problems, anticipated shifts in relative power, and the efficacy of available weapons technologies (Reiter 2003).

open up the possibility of a "costly peace" dynamic in the following sense (Powell 1993, Fearon 2012, Coe 2015). If the effort required to compete became large enough, a state might try to end the race through implementing despite the risks. Whether this dynamic is plausible likely depends on two sets of factors. One is whether large shares of states' resources can be used efficiently to improve their positions in the race. Some great powers might find it difficult to reallocate substantial societal resources to a research race.

Another factor is whether actors have palatable alternatives to engaging in the research competition. This relates to what we have called "enmity," but which is equally a measure of the degree of desperation that actors have to win the race. In most cases, rather than incurring unsustainably high costs in pursuit of a particular technology, we expect states to prefer investing in other technologies or alternative sources of strategic advantage. Of course, some states may be willing to invest heavily in a research race when a new weapons technology is seen as essential to security, although it is difficult to identify a single historical case where unsustainably high costs of competition resulted from investments in *research*. More commonly, the costs of arms *production* and the maintenance of standing armies have been seen as taxing on societies, as in the cases of the transitions to metal clad warships in the 19th century or nuclear weapons in the 20th. On the whole, therefore, we expect it to be the rare case that costly peace dynamics that result from effort levels have a significant influence on the overall level of risk in research competitions. The infungibility of societal resources means it is usually not effective to approach unsustainably high costs of research competition. Nevertheless, for potentially important exceptional cases, this represents another avenue for future research.

Do our results imply that middle powers should try to widen or narrow technological gaps between great powers by aiding one and/or hindering its rival? Suppose, for example, animosity between the US and China grows but one has a lead in some autonomous weapons technology, so that the laggard is willing to gamble on deploying its most advanced weapon before figuring out how to reduce its risks. The model suggests that the world might be

safer if other countries got involved, whether to help the laggard catch up (so the former laggard could compete without being reckless) or help the leader get even further ahead (so the laggard would give up competing at all).

Nonetheless, we caution against an overly general interpretation of these results. Trying to shrink, maintain, or increase the gap between competing great powers may not be a first-best option even from the perspective of third parties. It may be more effective to try to reduce enmity between rivals by creating institutions that facilitate trust or to use some combination of carrots and sticks to dissuade both sides from trying to deploy a particularly risky technology. In our view, the implication of the analysis is rather to identify dangerous moments in technology races and to suggest that the full range of policy options be considered in those cases.

Related to these questions about the desirability of close versus unequal races are the effects of what may be an ever more open world as offensive cyber capabilities outpace defensive capabilities.[61] If secrets become a thing of the past, at least between great powers, this will have a range of implications for technology races and technology implementation. When openness results in two actors with high enmity being at the same knowledge level, they can be expected to take on less risk than they would if one was further behind. Yet, openness will also heighten competitive dynamics by allowing incapable laggards to become capable, potentially increasing risk dramatically. We suspect that adding additional actors to a race will also increase risk in such a context. Where two neck-and-neck powers may refrain from dangerous implementation because each can hope to win the race to a technology that is both safe and effective, this prospect dims with more competitors. If openness becomes the *expectation*, this could also have other dangerous effects. Actors with a temporary advantage will have incentive to exploit it by taking greater risks in the short term. All of these dynamics merit further study.

[61] Saltzman 2013 argues that cyber capabilities are offense-biased, while Slayton 2017 argues that they are defense-biased. Garfinkel and Dafoe 2019 reconcile these theories by developing a model with an inverse-U relationship between investment in cyber technology and the offense-defense balance.

The model also provides considerations for scientists and engineers about when to pursue research aimed at improving the safety of the most powerful weapons technologies, as well as when to share the results of such research with rival powers. Marginal improvements in the safety of the riskiest weapons technologies can be counterproductive if they encourage states to deploy them when they would not have otherwise. For example, major military powers have refrained from fielding fully autonomous robotic weapons, out of fear that not having a "human in the loop" could result in horrific mistakes, including civilian casualties or accidental escalation in crises.[62] However, some experts and government officials have argued that over time, automated weapons will become increasingly safe and eventually less error-prone than humans.[63] Before we reach that point however, partial advances in safety may lead governments to decide that AI weapons are safe enough to deploy, even if substantial risks remain.[64]

These are complex and consequential dilemmas, and we do not mean to suggest that carrying out and sharing research aimed at improving the safety of weapons technologies is always harmful. As our results show, in situations of high enmity, states may already be willing to deploy their most powerful weapons: making those technologies safer in such cases reduces shared risks. However, researchers should consider the possibility that sharing incremental safety insights could backfire in some situations because of risk compensation. Our model suggests it would sometimes be preferable to wait to share the results of safety research until it has progressed far enough to substantially reduce the risk of powerful weapons technologies. In some other cases, researchers might do more good by prioritizing safety research on a less potent version of a technology, to encourage states to substitute it for a more

[62] Horowitz 2018; Future of Life Institute 2021

[63] See for example: Arkin 2010; Geist and Lohn 2018; U.S. Government 2018.

[64] Even in matters of nuclear warfare, the U.S., Russia, and China seem to be taking steps toward greater automation, such as systems for recommending retaliatory nuclear strike options for decision makers to consider (Klare 2020). Lowther and McGiffin (2019) argue that the US should consider creating a fully automated system to detect a nuclear attack and launch a retaliatory nuclear strike. Ideas for improving the reliability of automated nuclear command and control systems may therefore lead countries to take greater risks overall.

capable but riskier system.[65]

Conclusion

One theme in this analysis of risky technology races is the significance of actors who are not at the technological vanguard. The capabilities and intentions of these actors have substantial influence on the incentives of technology leaders. When laggards are capable, they are also dangerous: they have more incentive to trade safety for competitive advantage and when they are determined to be first, to influence events or to obtain status recognition, it is a trade they will make. States that fear they are falling behind technologically are more willing to base their security on capabilities and strategies that imply higher risk of escalation to great-power conflicts that would be catastrophic for all.[66]

Another theme is the surprising effects of safety research. We would expect safety research to improve welfare, but this is not always the case. Safety advances can embolden actors to risk implementation of a technology they previously considered too risky. Further, the bar for how effective a safety advance must be to increase welfare is higher the lower enmity is because actors with low enmity are already more cautious.

There is much more to be said about why decision makers make the technological risk calculations they do. We encourage more research in this area, and we believe it will return to the sentiments of Rachel Carson: "All this has been risked — for what? Future historians may well be amazed by our distorted sense of proportion."[67] It is essential to understand where our risk tolerances present the most danger to welfare. One reason is that the scope of potential catastrophes is probably increasing.

[65] Policymakers too should consider potential risk compensation before sharing safety insights about weapons technologies with other governments.

[66] Russia's efforts to develop tactical nuclear weapons capabilities to a greater extent than other countries may be an example of such a high-risk strategy. See Congressional Research Service (2021), *Nonstrategic Nuclear Weapons*.

[67] Carson 2002, 8.

References

Allen, Greg and Daniel Chan. 2017. Artificial Intelligence and National Security. Technical report Belfer Center for Science and International Affairs.
URL: *www.belfercenter.org/publication/artificial-intelligence-and-national-security*

Altmann, Jürgen and Frank Sauer. 2017. "Autonomous Weapon Systems and Strategic Stability." *Survival* 59(5):117–142.

Arkin, Ronald C. 2010. "The Case for Ethical Autonomy in Unmanned Systems." *Journal of Military Ethics* 9(4):332–341.

Axelrod, Robert. 1984. *The Evolution of Cooperation.* New York: Basic Books.

Ayoub, Kareem and Kenneth Payne. 2016. "Strategy in the Age of Artificial Intelligence." *Journal of Strategic Studies* 39(5-6):793–819.

Bas, Muhammet A. and Andrew J. Coe. 2016. "A Dynamic Theory of Nuclear Proliferation and Preventive War." *International Organization* 70(4):655–685.

Berger, Suzanne. 2000. "Globalization and politics." *Annual Review of Political Science* 3(1):43–62.

Bimpikis, Kostas, Shayan Ehsani and Mohamed Mostagir. 2019. "Designing dynamic contests." *Operations Research* 67(2):339–356.

Bommasani, Rishi, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong,

Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou and Percy Liang. 2021. "On the Opportunities and Risks of Foundation Models.".

URL: *https://arxiv.org/abs/2108.07258*

Borrie, John. 2019. Cold war lessons for automation in nuclear weapon systems. In *The Impact of Artificial Intelligence on Strategic Stability and Nuclear Risk: Volume I Euro-Atlantic Perspectives*, ed. Vincent Boulanin. Stockholm International Peace Research Institute pp. 41–52.

Bostrom, Nick. 2002. "Existential risks: analyzing human extinction scenarios and related hazards." *Journal of Evolution and Technology* 9.

Bostrom, Nick and Milan M. Cirkovic. 2011. *Global Catastrophic Risks*. Illustrated edition ed. Oxford: Oxford University Press.

Budd, Christopher, Christopher Harris and John Vickers. 1993. "A model of the evolution of

duopoly: Does the asymmetry between firms tend to increase or decrease?" *The Review of Economic Studies* 60(3):543–573.

Cai, Hongbin and Daniel Treisman. 2005. "Does competition for capital discipline governments? Decentralization, globalization, and public policy." *American Economic Review* 95(3):817–830.

Carson, Rachel. 2002. *Silent spring*. Houghton Mifflin Harcourt.

Coe, Andrew J. 2015. "The modern economic peace." *Unpublished manuscript, University of Southern California* .
URL: *datascience.iq.harvard.edu/files/pegroup/files/coe2017.pdf*

Danzig, Richard. 2018. "Technology Roulette: Managing Loss of Control as Many Militaries Pursue Technological Superiority.".
URL: *www.cnas.org/publications/reports/technology-roulette*

Denicolo, Vincenzo. 2000. "Two-Stage Patent Races and Patent Policy." *The RAND Journal of Economics* 31(3):488–501.

Ding, Jeffrey. 2018. "Deciphering China's AI Dream: The context, components, capabilities, and consequences of China's strategy to lead the world in AI.".
URL: *www.fhi.ox.ac.uk/deciphering-chinas-ai-dream*

Downs, George W and David M Rocke. 1990. *Tacit bargaining, arms races, and arms control*. Univ of Michigan Pr.

Farquhar, Sebastian, John Halstead, Owen Cotton-Barratt, Stefan Schubert, Haydn Belfield and Andrew Snyder-Beattie. 2017. "Existential Risk: Diplomacy and Governance.".

Fearon, James D. 2011. "Arming and Arms Races." *Annual Meetings of the American Political Science Association, Washington, DC* .

Fearon, James D. 2012. "A Simple Political Economy of Relations among Democracies and Autocracies." p. 34.

Fearon, James D. 2018. "Cooperation, conflict, and the costs of Anarchy." *International Organization* 72(3):523–559.

Future of Life Institute. 2021. "10 Reasons Why Autonomous Weapons Must Be Stopped.". **URL:** *futureoflife.org/2021/11/27/10-reasons-why-autonomous-weapons-must-be-stopped*

Garfinkel, Ben and Allan Dafoe. 2019. "How does the offense-defense balance scale?" *Journal of Strategic Studies* 42(6):736–763.

Geist, Edward and Andrew J. Lohn. 2018. How Might Artificial Intelligence Affect the Risk of Nuclear War? Technical report RAND.
URL: *www.rand.org/pubs/perspectives/PE296.html*

Goertz, Gary and Paul F. Diehl. 1995. "Taking "enduring" out of enduring rivalry: The rivalry approach to war and peace." *International Interactions* 21(3):291–308.

Hensel, Paul R., Gary Goertz and Paul F. Diehl. 2000. "The Democratic Peace and Rivalries." *The Journal of Politics* 62(4):1173–1188.

Hörner, Johannes. 2004. "A perpetual race to stay ahead." *The Review of Economic Studies* 71(4):1065–1088.

Horowitz, Michael C. 2018. "Artificial intelligence, international competition, and the balance of power." *Texas National Security Review* 1(3):36–57.

Horowitz, Michael C. 2019. "When speed kills: Lethal autonomous weapon systems, deterrence and stability." *Journal of Strategic Studies* 42(6):764–788.

Intriligator, Michael D. and Dagobert L. Brito. 1985. "Non-Armageddon solutions to the arms race." *Contemporary Security Policy* 6(1):41–57.

Jackson, Matthew O and Massimo Morelli. 2008. "Strategic militarization, deterrence and wars." *Deterrence and Wars (September 1, 2008)* .

Kahler, Miles. 1998. "Modeling races to the bottom." *Unpublished manuscript. http://irpshome. ucsd. edu/faculty/mkahler/RaceBott.pdf* .

Kania, Elsa B. 2020. "AI Weapons" in Chinese Military Innovation. Technical report Brookings.

URL: *www.brookings.edu/wp-content/uploads/2020/04/FP*$_2$0200427$_a i_w eapons_k ania.pdf*

Klare, Michael T. 2020. "'Skynet' Revisited: The Dangerous Allure of Nuclear Command Automation." *Arms Control Today* .

URL: *www.armscontrol.org/act/2020-04/features/skynet-revisited-dangerous-allure-nuclear-command-automation*

Kristensen, Hans M. and Matthew McKinzie. 2012. Reducing Alert Rates of Nuclear Weapons. Technical report United Nations Institute for Disarmament Research.

URL: *www.unidir.org/publication/reducing-alert-rates-nuclear-weapons*

Langinier, Corinne and GianCarlo Moschini. 2002. "The Economics of Patents: An Overview." *CARD Working Papers* 335.

URL: *www.card.iastate.edu/products/publications/pdf/02wp293.pdf*

Lowther, Adam and Curtis McGiffin. 2019. "America Needs A "Dead Hand"." *War on the Rocks* .

URL: *warontherocks.com/2019/08/america-needs-a-dead-hand/*

McKinnon, John D. and Asa Fitch. 2022. "U.S. Restricts Semiconductor Exports in Bid to Slow China's Military Advance." *The Wall Street Journal* . October 7, 2022 https://www.wsj.com/articles/u-s-restricts-semiconductor-exports-in-bid-to-slow-chinas-military-advance-11665155700?mod=article_inline (accessed November 10, 2022).

Moore Geist, Edward. 2016. "It's already too late to stop the AI arms race–We must manage it instead." *Bulletin of the Atomic Scientists* 72(5):318–321.

Murphy, Dale D et al. 2004. "The structure of regulatory competition: Corporations and public policies in a global economy." *OUP Catalogue* .

National Counterintelligence and Security Center. 2021. Protecting Critical and Emerging U.S. Technologies from Foreign Threats. Technical report.

Ord, Toby. 2020. *The Precipice: Existential Risk and the Future of Humanity*. Hachette Books.

Powell, Robert. 1993. "Guns, Butter and Anarchy." *American Political Science Review* 87(1):115–132.

Prakash, Aseem and Matthew Potoski. 2006. "Racing to the bottom? Trade, environmental governance, and ISO 14001." *American journal of political science* 50(2):350–364.

Price, Matthew, Stephen Walker and Will Wiley. 2018. "The Machine Beneath: Implications of Artificial Intelligence in Strategic Decisionmaking." *PRISM* 7(4):92–105.

Reiter, Dan. 2003. "Exploring the Bargaining Model of War." *Perspectives on Politics* 1(1):27–43.

Reynolds, Jesse. 2015. "A critical examination of the climate engineering moral hazard and risk compensation concern." *The Anthropocene Review* 2(2):174–191.

Rockett, Katharine. 2010. Chapter 7 - Property Rights and Invention. In *Handbook of the Economics of Innovation*, ed. Bronwyn H. Hall and Nathan Rosenberg. Vol. 1 of *Handbook of The Economics of Innovation, Vol. 1* North-Holland pp. 315–380.

Rudner, Tim G. J. and Helen Toner. 2021. Key Concepts in AI Safety: An Overview. Technical report Center for Security and Emerging Technology.
URL: *cset.georgetown.edu/publication/key-concepts-in-ai-safety-an-overview/*

Saltzman, Ilai. 2013. "Cyber Posturing and the Offense-Defense Balance." *Contemporary Security Policy* 34(1):40–63.

Sanger, David E. 2022. "China Has Leapfrogged the U.S. in Key Technologies. Can a New Law Help?" *The New YOrk Times* . July 28, 2022 https://www.nytimes.com/2022/07/28/us/politics/us-china-semiconductors.html (accessed November 10, 2022).

Scharre, Paul. 2016. Autonomous Weapons and Operational Risk. Technical report Center for a New American Security.
URL: *s3.us-east-1.amazonaws.com/files.cnas.org/documents/CNAS$_A$utonomous − weapons − operational − risk.pdf*

Schlosser, Eric. 2014. "Almost Everything in 'Dr. Strangelove' Was True." *New Yorker* .

Schmidt, Eric, Bob Work, Safra Catz, Mignon Clyburn, Steve Chien, Chris Darby, Kenneth Ford, José-Marie Griffiths, Eric Horvitz, Andrew Jassy, Gilman Louie, William Mark, Jason Matheny, Katarina McFarland and Andrew Moore. 2021. Final Report of the National Security Commission on Artificial Intelligence (AI). Technical report National Security Commission on Artificial Intelligence.
URL: *www.nscai.gov/2021-final-report*

Scotchmer, Suzanne and Jerry Green. 1990. "Novelty and Disclosure in Patent Law." *The RAND Journal of Economics* 21(1):131–146.

Slayton, Rebecca. 2017. "What Is the Cyber Offense-Defense Balance? Conceptions, Causes, and Assessment." *International Security* 41(3):72–109.

State Council of China. 2017. "Notice of the State Council Issuing the New Generation of Artificial Intelligence Development Plan.".

Swanson, Ana. 2022. "Congress Is Giving Billions to the Chip Industry. Strings Are At-

tached." *The New YOrk Times* . August 3, 2022 https://www.nytimes.com/2022/08/03/business/economy/chip-industry-congress.html (accessed November 10, 2022).

Thomas, M. A. 2020. "Time for a Counter-AI Strategy." *Strategic Studies Quarterly* 14(1):3–8.

U.S. Government. 2018. "Humanitarian Benefits of Emerging Technologies in the Area of Lethal Autonomous Weapons.".

von Braun, Joachim, Margaret S. Archer, Gregory M. Reichberg and Marcelo Sánchez Sorondo. 2021. Robotics, AI, and Humanity: Opportunities, Risks, and Implications for Ethics and Policy. In *Robotics, AI, and Humanity: Science, Ethics, and Policy*, ed. Joachim von Braun, Margaret S. Archer, Gregory M. Reichberg and Marcelo Sánchez Sorondo. Springer pp. 1–13.

Zhong, Raymond and Cao Li. 2022. "With Money, and Waste, China Fights for Chip Independence." *The New YOrk Times* . December 24, 2020. https://www.nytimes.com/2020/12/24/technology/china-semiconductors.html (accessed November 10, 2022).

Zwetsloot, Remco and Allan Dafoe. 2019. "Thinking about risks from AI: Accidents, misuse and structure." *Lawfare* .
URL: *www.lawfareblog.com/thinking-about-risks-ai-accidents-misuse-and-structure*