Uncertainty, information, and risk in international technology races

Nicholas Emery-Xu[*], Andrew Park[†], Robert Trager[‡]

August 26, 2023

Abstract

A formal model reveals how the information environment affects international races to implement a powerful, dangerous new military technology, which may cause a "disaster" affecting all states. States implementing the technology face a tradeoff between the safety of the technology and performance in the race. States face unknown, private, and public information about capabilities. More decisive races, in which small performance leads produce larger probabilities of victory, are usually more dangerous. In addition, revealing information about rivals' capabilities has two opposing effects on risk: states discover either that they are far apart in capability and compete less or that they are close in capability and drastically reduce safety to win. Therefore, the public information scenario is less risky than the private information scenario except under high decisiveness. Finally, regardless of information, the larger the eventual loser's impact on safety relative to the eventual winner's, the more dangerous is the race.

[*] UCLA, Department of Economics

[†] Uber

[‡] UCLA, Department of Political Science. We would like to thank Paolo Bova, Richard Jordan, Lennart Stern, Jeffrey Ohl, and Jonas Mueller, as well as conference participants at APSA 2021, Conference on Great Power Competition in the 21st Century (2022), and the Politics of Emerging Technology Mini-Conference (2022), for helpful feedback and comments. We thank the Stanford Existential Risks Initiative and the Open Philanthropy Project for financial and logistical support. All remaining errors are our own.

1. Introduction

Uncertainty is central to the study of arms races in the field of international relations. It underpins analyses of when arms races lead to conflict [Kydd, 1997, Jervis, 1976, Schelling, 1980] and studies of the potential of treaties and other forms of international cooperation [Kydd and Straus, 2013]. In a world characterized by anarchy, it is likely that the information environment will play a key role in determining the impact of emerging technologies and the competitions to develop them. We study the role of incomplete information in a setting that has been largely neglected in the international relations literature: races for powerful new technologies. Some scholars posit that races for such technologies may become a key feature of international politics in the coming decades, as states compete to be the first to develop new technologies such as advanced artificial intelligence (AI) or nanotechnology that could give them a sudden, significant increase in capability over other states.

Such races have important differences from competitions to build larger numbers of existing armaments.[1] An important feature of these races is that they are associated with different kinds of risk. Sometimes this risk is an exogenous feature of the international system that is exacerbated by an arms race. If a state's technological development has the potential to cause a relative power shift, its rivals may attack to prevent such development [Fearon, 1995]. In other cases, which we focus on in the present work, the risk of negative externalities is inherent to the development and implementation process itself. For example, biological weapons development can lead to releases of pathogens that affect a broad range of actors beyond those involved in development. In April 1982, a research lab that was part of the Soviet biological weapons program produced an anthrax outbreak in the city of

Sverdlovsk that killed over 100 people. Likewise, genetic sequencing of the virus that caused the 1977-1978 influenza epidemic reveals that the virus seems likely to have come from a research laboratory [Rozo and Gronvall, 2015]. Biological weapons use has a relatively high probability of infecting the user, an argument put forth about why there have been relatively few uses of such weapons despite the weakness of the provisions of the Biological Weapons Convention [Ord, 2020]. Indeed, some scholars posit that substantial global risks many result from such technology races [Cave and ÓhÉigeartaigh, 2018, Stern, 2002]. Because of these risks, actors face an inherent *safety-performance tradeoff*, in which they must choose the optimal allocation of resources between advancing the performance level of a technology, thereby increasing the probability of winning the race, and investing in the safety of the technology, which lowers the risk of disaster [Trager et al., 2021]. Such allocations are determined by the strategic contexts in which the actors find themselves. Actors' information about their rivals' capability interacts with this tradeoff in a number of ways. Overestimation of a rival's technological capability may lead an actor to overinvest in capability, increasing risk relative to the complete information scenario [Stafford et al., 2021]. In other cases, actors may learn that they are far behind in the race and choose to cede the prize to their opponent, lowering risk [Bimpikis et al., 2019].

Some of these dynamics are illustrated in the race for the first nuclear bomb during World War II when physicists involved in the Manhattan Project expressed concerns over the safety-performance tradeoff. Edward Teller, for example, feared that a nuclear fusion reaction could ignite the atmosphere, ending life on Earth. He privately urged the U.S. government to delay development so that additional calculations and tests could be performed. Though the team was able to show that these fears were improbable, Teller and his colleagues remained worried until after the Trinity test was conducted. Part of the reason why history favored development over safety is the U.S. government's uncertainty over the level of progress of Germany's development of nuclear weapons. Albert Einstein, for example, would later write to US President Franklin Roosevelt, "Had I known that the Germans would not succeed in developing an atomic bomb, I would have done nothing for the bomb." [Newsweek, 1947].[2]

A number of scholars believe that the development of advanced forms of artificial intelligence will exhibit similar strategic dynamics [Russell, 2019, Yudkowsky et al., 2008]. Indeed, AI has already begun to shape international and domestic politics in profound ways, expanding the use of fully automated drones [Horowitz, 2018] and heightening domestic surveillance in authoritarian regimes [Beraja et al., 2023]. And progress in the field is increasing rapidly. On average, experts in the field believe there is a 50% chance of developing an AI that surpasses human performance on all job tasks by 2060 [Zhang et al., 2022]. Such advanced AI systems may carry enormous benefits to states, but they all produce distinct risks, ranging from AI objectives that are misaligned with human incentives or control of powerful AI systems by expansionary states or other dangerous actors [Russell, 2019, Brundage et al., 2018].[3] Though AI development is currently led by scientists a strong value for open sharing of knowledge, as the rewards from advanced AI become more apparent, as with nuclear weapons, states may have incentives to increase the pace and secrecy of development with as-yet-unknown effects on risk.

Likewise, the importance of such a tradeoff is likely to become increasingly important in biological research for both military and civilian use. The field of synthetic biology has an explicit goal of reducing the level of tacit knolwedge necessary to produce new biological agents [Mukunda et al., 2009], which could increase the ability of state and non-state actors to access the technology. The proliferation of terrorist actors concurrent with the publication of gene sequencing for Ebola, influenza, and other deadly pathogens on the Internet led to an increased focus on preventing proliferation [Stern, 2002]. Likewise, the cost of genome sequencing, and thus of developing deadly biological agents, has been halving faster than every two years, making the development of weaponized agents accessible to an ever-increasing set of actors [Mukunda et al., 2009]. Finally, accidental laboratory leaks can also cause deadly outbreaks, exacerbating risk [Rozo and Gronvall, 2015, Lipsitch and Inglesby, 2014]. Such a rapidly evolving threat has forced political actors to consider a safety-performance tradeoff for biological research. Stern [2002], for example, argues that restricting the development of basic biological research could slow diffusion but also hinder scientific innovation.

We develop a formal model that captures many of these strategic considerations. We solve for the perfect Bayesian equilibria under three scenarios regarding information about capabilities: unknown, private, and public. First, we show that more decisive races, in which small leads in performance produce larger probabilities of victory in the race, are weakly more dangerous under most parameter values. Second, we show that revealing information about the capabilities of rivals has two opposing effects on disaster risk. The benefit is that actors may discover that they are sufficiently far apart in capability and will compete less. The cost is that actors may discover they are close in capability and thus engage in a dangerous race to the bottom, cutting corners on safety to win the race. As a result, the public information scenario is more dangerous than the private information scenario only under high decisiveness. As decisiveness decreases, the first effect dominates the second, so that public knowledge of capabilities is welfare-improving. Third, in all information scenarios, we find that the larger the impact of the eventual loser on safety, relative to the eventual winner, the more dangerous is the race due to a public-good effect.

Our work is organized as follows. Section 2 provides an overview of the interaction among information, investments, and risk in the arms race literature, finding that existing models fail to fully capture the strategic situation in which states find themselves when developing risky new technologies. Section 3 describes our choice of model primitives, grounded in existing cases of technology races. Section 4 presents the base model. Section 5 describes the forces that generate risk under each information scenario and illustrates these forces in a series of historical examples. In Section 6, we consider the role of safety sharing, enmity, and regime type on risk. Finally, Section 7 concludes.

2. Information, arms racing, and risk

An extensive literature exists on how information affects the risk of conflict in arms races. The sorts of incomplete information that drive the risk of conflict appears to fall into three broad categories [Ramsay, 2017]. The majority of the literature has focused on uncertainty over actors' costs of conflict [Kydd, 1997]. A second strand of literature focuses on psychological factors influencing states' risk-taking, invoking such causes as states' mutual tendency to be either overly optimistic about their own chances of winning a conflict [Wittman, 2009] or overly pessimistic about the intent of a rival's arms buildup [Jervis, 1976]. Finally, a third strand of literature, in which our work is situated, focuses on the role of uncertainty about the capabilities of rivals. Across literatures, the existence of a baseline bargaining model of conflict [Fearon, 1995] has given scholars a framework with which to analyze the role of uncertainty in war. This has led to a number of robust analytical results, including that weaker types are less likely to initiate conflict [Powell, 2004], that a higher variance over the distribution of types increases risk [Reed, 2003, Wittman, 2009], and that perfectly peaceful equilibria only obtain when the joint cost of war is large enough [Fey and Ramsay, 2011].

Existing literature has studied the role of information in *quantitative* arms races, those for which states accumulate arms but the level of technology remains fixed.[4] Kydd [2000] and Meirowitz and Sartori [2008] focus on situations in which states are able to arm in private before bargaining. Kydd shows that states perceived as having relatively low capabilities tend to arm in private in order to secure better bargaining outcomes. Meirowitz and Sartori [2008] endogenize the decision to disclose capabilities, arguing that states prefer to keep their capabilities private to secure better bargaining outcomes, even when the risk of war increases. A second class of models studies an asymmetric arms race, when a weaker state is seeking to acquire new military capabilities to lower the gap with strong states. Bas and Coe [2016] study a dynamic model in which a strong state obtains a noisy signal about an arming state's level of capabilities, finding that the estimated time to completion of the arming is more predictive of preventative war than the mere existence of arming.[5]

The empirical evidence on the influence of arms races on war is mixed. Early work by Richardson [1960] and Wallace [1979, 1982] finds a positive correlation between rapid accumulation of arms and the outbreak of war. Later work has qualified these results, finding only some types of arms races are correlated with an increased outbreak of war. Horn [1987] finds that brief periods of rapid arming do not heighten the risk of war, while Sample [1997] finds that, while most arms races are associated with an increased probability of conflict, those involving nuclear weapons buildups or those between bitter rivals are not.

In contrast to the study of arms races and war, the study of *qualitative* arms races has been hindered by the lack of a standard model for thinking about such competitions. As such, existing formalizations of arms races we argue are poor descriptors of the strategic environment in which states find themselves. First, the rewards to qualitative races may be inordinately large, perhaps leading to rapid, discontinuous power shifts between a state that develops a new technology and her geopolitical rivals. This view is epitomized by President Vladimir Putin of Russia, who said with regards to military uses of AI: "the one who becomes the leader in this sphere will be the ruler of the world." [AP, 2017]. As such, it makes little sense to view the outcome of the race in terms of bargaining over shares of a pie [Fearon, 1995]. Second, the risks resulting from such races may be quite different along a number of dimensions compared to quantitative arms races, whose main risks are war and an inefficient allocation of economic resources.[6] Instead, risks from technology races may be both much larger in consequence and have a far smaller probability of realization than risks from quantitative races [Stern, 2002]. However, states have a far greater control over the level of risk than they do over the occurrence of war.[7] For example, biological research can continue, albeit at a slower pace, even if a state were to prohibit the publication of most biological research and keep risk to a minimum.[8] This allows states to control the tradeoff between risks from speeding up development and the risk of a geopolitical rival developing the technology first.

In this work, we develop a model that explicitly takes into account such features. This model contributes to a small but growing literature on qualitative arms races. Naude and Dimitri [2020] study an evolutionary model of a qualitative race within one country, showing that taxing technological development and using public procurement can incentivize cooperation and reduce risk. Stafford et al. [2021] employ a dynamic model to study the effects of a capabilities gap on risks in qualitative races, finding that risk is higher for a larger gap in players' performance levels when enmity is high but is lower when enmity is low. They show that there exists a safety-performance tradeoff in which investments in safety and investments in research progress are complementary goods. Both of these models assume that research capability levels are common knowledge, an assumption that seems unlikely in real-world races. Both the Manhattan Project and the Soviet bioweapons program, for example, were carried out under conditions of high secrecy. While the model in Armstrong et al. [2016] compares risk in a technology race under public and private knowledge of capabilities, they make the strong assumption that states have perfect knowledge of the R&D process such that the state with the highest performance wins the contest for certain. In contrast, we generalize their model to take into account both the role of information about research progress *and* the capabilities of rivals. It is to this uncertainty that we now turn.

3. Model primitives

In our model, two states $i \in \{1, 2\}$ compete to build a significant military technology, such as a new biological weapon or powerful AI system. Each state is endowed by nature with a research capability level x_i, which we can think of as determined exogenously by existing attributes such as GDP or military expenditures.[9] Depending on the information state, explained in greater detail below, x_i may be unknown, privately known, or publicly known. Each state's capability is drawn independently from a commonly-known distribution $G(x_i)$, which we assume for simplicity and without much loss of generality, is uniformly distributed on $[0, \mu]$. Each state chooses a level of safety investment $s_i \in [0, 1]$. However, investing in safety research detracts from a state's performance in the race in a linear fashion. We denote a state's net performance level as $k_i = x_i - s_i$. If a state i wins the race, it then implements the technology and receives its military benefits; with probability s_i, implementation is successful, and with probability $1 - s_i$, a disaster is incurred.[10] For now, we assume that only the winner has a chance to implement the technology and thus contribute to the safety of the process, though later we relax this assumption and allow both states to affect the risk of disaster. We normalize the value of winning the race to 1 and the value of a disaster to 0. If i loses the race, its rival j has a chance to implement the technology. We assume that the disaster affects all states equally.[11] However, if a rival wins the race and implementation is successful, a state receives only an intermediate payoff $(1 - \eta)$, where $\eta \in (0, 1]$ represents enmity, or the opportunity cost of losing the race. In effect, then, we have the following utility ordering for each state i:

$$\mathbb{E}[u_i|(\text{i implements})] > \mathbb{E}[u_i|\text{j implements}] \geqslant \mathbb{E}[u_i|\text{disaster occurs})]$$

where $u_i(.)$ is given by

$$\mathbb{E}[u_i(s_i)] = s_i Pr(i \text{ wins}|k_i, k_j) + (1 - \eta)s_j Pr(j \text{ wins}|k_i, k_j)$$

Finally, we assume that states' performance levels k_i translate into success in the race according to a logistic contest success function (CSF):

$$Pr(i \text{ wins}|k_i, k_j) = \frac{\mathrm{e}^{mk_i}}{\mathrm{e}^{mk_i} + \mathrm{e}^{mk_j}}$$

Though contest success functions have been commonly used to model war outcomes (cf. Fearon [2018], Skaperdas [1998], Hirshleifer [1995]), their use in a technology race warrants discussion. First, the use of a CSF allows us to model the level of uncertainty inherent in innovation. Even given a known performance level, research outcomes are the result of an inherently random process that involves a certain amount of luck as researchers seek to recombine existing knowledge in novel ways [Weitzman, 1998].[12] Even if another research team is lower in capabilities, they have a positive probability of making the discovery first. Thus, it is common for economists to measure innovation races between firms using a contest success function [Baye and Hoppe, 2003]. Second, we choose the logistic CSF in particular because degree of *difference* in capabilities matters for success. This appears to be an important feature in determining risks from qualitative races. Both the Manhattan Project scientists and Soviet biologists stated that the potential of their rival, Nazi Germany and the United States, respectively, to catch up in the race was a driving factor encouraging them to favor performance over safety [Ord, 2022].[13]

An important focus of our paper is the decisiveness parameter in the CSF, $m \geqslant 0$. This determines the rate at which additional effort translates into success. At low values of decisiveness, progress is highly uncertain, such that even a state with relatively low capability may win the race due to “luck” or some other resource that is not accounted for by x, but as $m \to \infty$, the state with the highest value of k wins the race with certainty.[14] In our context, m is correlated with the level of uncertainty over the research process.[15] In cases when m is low, mastering the steps necessary to build a technology is likely to be difficult, requiring high levels of expertise or a search through a wide space of ideas in order to find one that "works." In other cases, m is high, representing a high level of clarity about how research capabilities translate into success. This is likely to be the case if a new technology is an iteration of an existing one. Summarizing the above description, Figure 1 presents the structure of the game.

4. Base model

To examine the risk that arises under different information structures, we begin by characterizing the unique symmetric Bayesian Nash Equilibrium of the game under no information, private information, and public information conditions. This will allow us to calculate the expected risk of the race given the distribution of states' capabilities $G(x)$.[16]

4.1 No information

We begin with the no information case. In this scenario, no state knows its own capability. This is distinct from the uncertainty that comes from decisiveness. With low values of decisiveness, states may know their own performance level and can channel their resources toward developing the technology, with uncertain results. Here, states also have no information about their own capability. This is a more fundamental source of uncertainty: does a state's stock of resources even contribute to technological progress at all? Realistically, then, the no information case represents a lower bound on states' knowledge, as in the real world states are likely to have at least an understanding of how to build a novel technology. Because states have the same prior beliefs over the type space, in the symmetric Nash equilibrium, each will choose the same strategy. Here we derive the equilibrium safety level of each player as well as the expected disaster risk over the distribution of states' capabilities.

Proposition 1. *In the case in which states do not know their capabilities, the unique symmetric BNE strategy is given by* $s^*_\varnothing = \min\left\{1, \frac{\mu}{2\eta[F(\mu)-F(0)]}\right\}$.

First, we note that the equilibrium outcome is written in terms of $F(c)$, where $C_i = X_i + V_i - V_j, V \sim \mathit{Gumbel}(1, \frac{1}{m})$ represents the capability level of state i, adjusted for uncertainty in winning, which we recall is parameterized by the race decisiveness m.[17] Here, we see that safety is inversely related to the number of states in the race, and the enmity level, while it is positively related to both the variance of the true types, parameterized by μ, and the variance of the noise (since as m decreases, the difference $F(\mu) - F(0)$ shrinks to zero). Note that since states cannot condition on their research capabilities, all play the same equilibrium safety level.

Now we turn to the disaster risk. This is the expected probability of disaster over the distribution of types $G(x)$ who play their BNE strategies. Since all states are playing the same action, the expected risk of disaster is simply given by $1 - s^*_\varnothing$.

Corollary 1.1. *In any distribution of capability levels the expected level of disaster risk is given by* $D_\varnothing = \max\left\{0, 1 - \frac{\mu}{2\eta[F(\mu)-F(0)]}\right\}$.

This value represents the expected risk from a qualitative race to a neutral third party who cannot query states for their types. This provides a useful benchmark by which we can compare the level of risk even as states are permitted increased knowledge of their own or their opponent's capabilities.

4.2 Private information

In this section, we consider the case in which each state knows its own capability level but not its rival's. This situation more closely resembles realistic qualitative races, which often involve closely-guarded state secrets. Power-seeking states have a strong incentive to keep their technological capabilities hidden from rivals in order to win the race, as even proof-of-concept demonstrations could lead to increased competition [Bimpikis et al., 2019]. The Soviet biological weapons program, for example, was conducted in so-called "closed cities," which were not printed on most maps and which were off-limits to even Soviet citizens. In the private information case, each state i can condition its safety level on its own capability, choosing a strategy $s_{private}(x_i)$. Before we establish the level of safety in this scenario, we begin with a lemma showing that performance, or capability less safety, is always increasing in x_i:

Lemma 1. *Let $k_i(x_i) = x_i - s_i(x_i)$. In the private information scenario, at any BNE, k_i is strictly increasing in x_i.*

Now we solve for the equilibrium safety level.

Proposition 2. *There exists a unique symmetric Bayesian Nash Equilibrium in pure strategies. The strategy is given by $s^*_{private}(x_i) = \min\left\{\frac{\int_{-\infty}^{x_i} F(c)^\eta dc}{F(x_i)^\eta}, 1\right\}$.*

Here, we see that states increase their safety levels as capabilities increase. Since more capable actors are more likely to win the race, knowing they are at the high end of the distribution, they can condition on this information by trading off additional performance for a higher level of safety. In addition, as in the private information case, we see that as enmity level increases, states start putting less efforts into safety. Finally, as before, the overall disaster risk is given by $D_{private} = 1 - \mathbb{E}_{winner}[s^*_{private}(x_i)]$, where $\mathbb{E}[.]$ is the expectation over the true distribution of player types. Corollary 2.1 gives formal expression of the disaster risk.

Corollary 2.1. *The disaster risk in the private information scenario is given by* $D_{private} = 1 - \frac{2}{\mu} \cdot \int_0^\mu \min\left\{\frac{\int_{-\infty}^x F(c)^\eta dc}{F(x)^\eta}, 1\right\} F(x)dx$.

4.3 Public information

Now we solve for the case in which both states are fully aware of each other's capabilities. While states are often incentivized to keep capabilities secret in order to impede rivals' progress, in other cases, states may want to demonstrate their capabilities to deter potential rivals, as in the case of the U.S. and Soviet hydrogen bomb tests in 1952 and 1953, respectively. Alternatively, this corresponds to states of the world in which espionage techniques make secret-keeping impossible. In this state of the world, the *difference* in capabilities determines states' safety choices and the risk of disaster. Here, we denote the leader's capability as x and the laggard's as y. Denote $\Delta := x - y$ as the variable on which states condition their safety choices. We can then find a unique pure strategy Nash equilibrium.

Proposition 3. *There exists a unique pure strategy Nash equilibrium for the public information cases for all values of* $m > 0$.

Denote the solution to this system of equations as $s^*(\Delta)$. Except in the case where

$m \to \infty$, equilibrium strategies do not permit a closed-form solution. In order to give intuition, then, in the following two corollaries, we show payoff and strategy equivalence with Armstrong et al. [2016] in the limit.

Corollary 3.1. *[Strategy equivalence] As $m \to \infty$, strategies converge to the following expressions:*

$$\lim_{m\to\infty} s_x^*(\Delta) = \min\left\{1, \tfrac{\Delta}{\eta}\right\}, \quad \lim_{m\to\infty} s_y^*(\Delta) = (1-\eta)\cdot\min\left\{1, \tfrac{\Delta}{\eta}\right\}.$$

Corollary 3.2. *[Payoff equivalence] As $m \to \infty$, states' utilities converge to the following expressions:*

$$\lim_{m\to\infty} u_x(\Delta) = \begin{cases} \frac{\Delta}{\eta} & \frac{\Delta}{\eta} < 1 \\ 1 & \textit{otherwise} \end{cases} \quad \textit{and} \quad \lim_{m\to\infty} u_y(\Delta) = \begin{cases} (1-\eta)\frac{\Delta}{\eta} & \frac{\Delta}{\eta} < 1 \\ 1-\eta & \textit{otherwise} \end{cases}.$$

From these limit expressions, we see that safety is positively associated with the difference between the leader and laggard and negatively associated with enmity. Unlike in the private information scenario, where highly-capable states always play higher levels of safety, here capable states will cut corners on safety if their rival is close in capability. As usual, enmity is negatively associated with safety. We compute the associated disaster risk as follows:

$$D_{public} = 1 - 2\int_0^{\mu}\int_0^{y}\Big[s_x^*(\Delta)F(x) + s_y^*(\Delta)(1-F(x))\Big]g(y)g(x)\,dx\,dy. \tag{1}$$

5. Information and risk

We now turn to comparisons of disaster risk under our three information scenarios.[18] We begin with a derivation of comparative statics in the model and then present empirical examples of these forces from real-world qualitative races.

5.1 Information and welfare

Changing information states can make the race more dangerous; we seek to understand how this interacts with the decisiveness parameter m. We present two primary sets of results, both illustrated in Figure 2. First, across most parameter values, the expected disaster risk is increasing with m. We prove strong versions of this statement for the no information case and private information case and a weaker statement for the public information case.[19]

Proposition 4. *In the no information and private information scenarios, risk always increases with decisiveness, unless risk is 0. In the public information scenario, risk is higher as* $m \to \infty$ *than as* $m \to 0$.

That is, if trading off safety for additional performance does not produce a high chance of winning the race, states will be reluctant to do so. As decisiveness becomes arbitrarily large, however, even the smallest additional unit of performance will determine for certain who wins the race, offering states a large incentive to cut corners. An observable implication of this is that in early stages of development, when progress is highly uncertain, states are likely to prioritize safe development over winning the race. In the early stages of nuclear weapons development, for example, and as late as 1942, U.S., Soviet, and German scientists published major insights on nuclear fission in publicly-accessible physics journals. Only around 1940, when the prospect of succeeding in the race became more imminent, did U.S. scientists begin concealing their results and progressing without checking necessary calculations [Ord, 2022].

Second, we want to know how the *relative* openness of the race interacts with decisiveness to influence risk. As shown in Figure 2, for much of the parameter space, the no information scenario is safer than the other two information scenarios. In fact, in Proposition 5 below, we show that the no information scenario is always at least as safe as the private information scenario.

The most interesting case, however, is the comparison between the public and private information scenarios. In quantitative arms race models, better information about rivals' capabilities tends to reduce risks [Wittman, 2009, Reed, 2003]. In contrast, in their study of a qualitative race, Armstrong et al. [2016] find that, given a high degree of enmity between states, the public information scenario is *more* risky than the private information scenario. However, we show that both effects are possible in a qualitative race. For large values of m and high enmity, we show that indeed public information produces higher risk than private information, producing an information hazard effect. However, as m declines, we see in Figure 2 that the public information scenario becomes *safer* than the private information scenario. Finally, as m tends to 0, we see that in both cases, players are unwilling to take any risk and implement at the maximum safety level. We present this result formally.

Proposition 5. *The no information scenario is always safer than the private information scenario, while the relative safety of the public and private information scenarios depends on m.*

These results are presented in Figure 2, where we see that the riskiest information scenario changes at $m = 6$. Two forces drive this result. Note that the decisiveness parameter m enters into the disaster risk function in two places: the contest success function and the equilibrium safety choices of the players. To see how these forces affect risk, consider the drivers of risk when $m \to \infty$. The public information scenario is riskier than the private information scenario as long as

$$\mu > \frac{(\eta+1)^3 + \eta^2}{3\eta} \tag{2}$$

In the public information case, risk is driven by cases in which the laggard is close behind the leader in capability, which happens with probability $G(x) = x/\mu$. In the private information case, risk is caused by low-capability winners, which, since there is no noise in the CSF, occurs precisely when *both* states have low capabilities, which occurs with probability $G(x)^2 = x^2/\mu^2$. Thus, the probability of a high-risk outcome decreases linearly in μ in the public information scenario but quadratically in μ in the private information scenario.[20] Thus, when $m = \infty$ and μ is sufficiently large, the public information case is most dangerous. Now fix μ and consider what happens as m tends to 0. In both cases, there is an increased probability that the laggard wins, which by itself increases overall risk. However, states also see a lower expected return to reducing safety investments, which implies lower risk. Consider what happens to the probability that i wins as m tends away from $+\infty$. We have

$$\frac{\partial}{\partial m} Pr(i \text{ wins}|x_i) = \frac{m e^{m(x_j - x_i + s_i - s_j)}}{(1 + e^{m(x_j - x_i + s_i - s_j)})^2} \tag{3}$$

In the public information case, $x_j - x_i + s_i - s_j \approx 0$, so $\partial P_i/\partial m \approx m/4$. Thus, the probability that a state wins is declining linearly in decisiveness. In the private information case, however, the expected gap in capabilities between the winner and loser at $m \to \infty$ is $1/6\mu$, so $\partial P_i/\partial m \approx me^{m[-6(\eta+2)/\mu(\eta+1)]}/(1 + e^{m[-6(\eta+2)/\mu(\eta+1)]})^2$, which note is close to 0 for m large. Thus, in the public information case, since the leader's probability of winning declines rapidly with m, we should expect that state and, in response, its rival to rapidly increase safety as decisiveness falls. In the private information scenario, the probability that a given state wins initially changes very little as decisiveness falls; thus, states will increase their safety strategies much more slowly than in the public information case. These effects are illustrated in Figure 3. In this figure, we simulate a race in which $\mu = 1, \eta = 0.9$. We fix $x_j = 0.5$ and consider what happens to expected safety when we vary x_i. In the public information case, the race is most dangerous when $x_i = x_j = 0.5$. As m falls, both leader and laggard rapidly increase safety, so the expected safety of the winner rises rapidly, even when $x_i = x_j$. In the private information case, risk is largely driven by competition with risky laggards. When $m \to \infty$, risk is driven entirely by the leader's safety choice, so safety is constant as long as $x_i \leqslant x_j$. As m falls to 5, we see that safety is still largely flat when $x_i \leqslant x_j$, indicating that the leader's efforts are mostly driving safety (since $s_i(x_i)$ is increasing in x_i). Since neither player has as much incentive to change strategies as m falls, the decline in the expected safety of the winner is not nearly as rapid as in the public information case. As a result, even if the public information scenario is more risky at high decisiveness, if decisiveness declines far enough, the private information case may well become more risky.[21]

Finally, the no information case is weakly safer than the private information case. Thus, learning one's own capability does not increase welfare. In the no information scenario, just as in cases with low decisiveness, states are quite uncertain about the returns to their own efforts. Here, the chance of being a laggard is exactly 1/2, so states are highly uncertain whether cutting corners on safety will benefit or harm them. Since they know their rivals will also face the same uncertainty, they will be quite unwilling to cut corners on safety. Thus, even for very high decisiveness, states are unwilling to take much risk, so states would be better off if they did not learn their type.

5.2 Empirical illustrations

Using our model, we now examine a series of examples that illustrate the importance of the above forces in historical technology races. We find that our analysis can explain the behavior of laggard states in a number of cases, illustrated in Table 1, which displays the relative level of risk taken on by the laggard in private and public information scenarios when the gap with a capable technology leader is varied.

5.2.1 Information scenarios

First, we compare the role of public and private knowledge when the gap between a capable leader and a laggard in a technology race is believed to be large. Consider the U.S.-Soviet race for the development of intercontinental ballistic missiles (ICBMs) in the 1950s. In November 1958, the U.S. tested its first successful ICBM. At the time, the U.S. Air Force estimated that the Soviet Union was far ahead in development, with hundreds of missiles developed as early as 1959 [Ellsberg, 2017]. Public statements by Soviet leadership supported this assessment: in October 1957, Premier Nikita Khrushchev declared Soviet factories were "turning out missiles like sausages" [Ellsberg, 2017]. As a result, the U.S. engaged in a rapid buildup of ICBM capabilities, producing over 1,000 by 1965 [Initiative, 2022]. The belief that they were relatively low on the distribution of capabilities, then, led the U.S. to pursue a risky development strategy, one which might not have occurred if information on capabilities was made public. While the Soviet Union did prioritize an ICBM program as early as 1957, Soviet leadership knew they were ahead in the capabilities distribution and proceeded more slowly with development, with Khrushchev acknowledging his statements about Soviet progress were a bluff [Mathers, 1998]. By 1961, when the U.S. learned that their estimates were mistaken, they had developed 40 missiles, while the Soviets had only 4 [Lawler and Mahan, 1961]. On the other hand, when states know they are behind in the race, they are often willing to develop safely. Consider the example of the Nuclear Non-Proliferation Treaty. That the weapons stockpiles of most nuclear powers are public information shows most states are willing to publicly reveal their nuclear capabilities in order to be able to develop safe, peaceful nuclear technology under the treaty in exchange for technology transfers, rather than engage in a weapons development race that risks strong international retaliation [Fuhrmann and Lupu, 2016]. A deviation to a race would be futile, since the wide gap in the distribution of capabilities is public knowledge. Thus, even most states that have technical capabilities to build nuclear weapons, such as Argentina and Brazil, are not willing to run the risk of a race [Narang, 2017]. Only when enmity is high enough are some low-capability states, such as North Korea, willing to begin active development.

Our model predicts that we should see the opposite effect when states are close: when information is public, states engage in a "race to the bottom" on development, generating a far higher level of risk than in the private information scenario. The U.S. government, for example, began the Space Race after the public launch of *Sputnik I* by the Soviet Union.[22]

Prior to the launch, while each state knew the other possessed a satellite program, the U.S. underestimated Soviet progress, with President Eisenhower choosing to deprioritize the speed of the U.S.'s own program. The Soviet launch of *Sputnik I* in October 1957 caught U.S. policymakers by surprise, prompting Eisenhower to re-prioritize the Vanguard project, which until then had been beset by delays [Barnhart, 2021]. After all, the U.S. was a close second, launching its first satellite a mere 4 months later, in January 1958. In contrast, by refusing to conduct public tests of a new technology, a state may be able to hide its level of technological prowess to avoid a costly race with its rivals. Consider the case of South Africa's nuclear weapons development. From 1971 through the dismantling in 1990-1991, South Africa pursued what Narang [2017] calls a "hidden" development strategy, prioritizing secrecy above speed or deterrence. Besides risks inherent in the production process, South Africa also faced risks of international isolation or retaliation should its program be discovered, as neither superpower wanted it to acquire them [Rabinowitz and Miller, 2015, Liberman, 2001]. While the U.S., suspicious of South Africa's refusal to accept International Atomic Energy Association inspections of its nuclear reactors, placed an embargo on nuclear fuel exports to South Africa, South Africa was able to assemble and then store its first thermonuclear weapon in 1979 and developed a further 5 by 1990 [Liberman, 2001]. Unlike the space race, South Africa's secret development did not prompt further proliferation by rivals, who did not know how behind they were in development, nor more than a moderate response by the international community, who did not realize the scale of South Africa's technological advance. In fact, a race was avoided despite a relatively high degree of enmity between South Africa's government and the two superpowers.

5.2.2 Decisiveness

The second primary source of risk in our model is high decisiveness. Though quantifying this parameter is difficult, we note that states' beliefs about m are likely to be highly correlated with how rapid, conditional on investment, they anticipate research progress to be. If states express a belief that research progress will happen quickly or that they are certain about key technological parameters, it is likely they believe m to be large. By analyzing variation in actor's expressions of certainty, then, we can analyze how changes in decisiveness affect states' safety choices. For example, as R&D on the first atomic bomb progressed, U.S. and British politicians and scientists increasingly expressed certainty that such a weapon could be developed in the near future. Early in development, however, there was considerable uncertainty about the size of the critical mass of uranium and amount of infrastructure spending that would be required to develop a bomb. In 1939-1940, estimates for the critical mass ranged from 44 tons to 1 pound of uranium-235, indicating high uncertainty about the timeline required to produce sufficient enriched uranium [Ord, 2022]. In this period, it could be said that decisiveness was low: it was unclear by how much a marginal investment in capabilities would improve performance in the race. At this time, scientists and officials seemed relatively unconcerned about sacrificing safety for an advantage in the arms race, as U.S., German, and Soviet scientists continued to publish key results on nuclear fission in physics journals [Ord, 2022]. By 1942, the U.S. and U.K. governments came to believe that decisiveness was considerably higher. For one, uncertainty surrounding estimates for the critical mass had declined to such a degree that the U.S. was able to accurately forecast the remaining time to the development of uranium and plutonium bombs [Ord, 2022]. As uncertainty in the possibility of success decreased, the U.S. drastically increased funding of capabilities: the Manhattan Project would end up totalling 0.4% of U.S. GDP by 1944. Likewise, U.S. funding on safety relative to performance began to decrease, as it pressed forward with the Trinity test despite Teller's fears that it could ignite Earth's atmosphere [Ord, 2020].

6. Multiple tests, varying enmity

Up until now, we have made a number of simplifying assumptions: that only the winner can implement the technology, that enmity is fixed, and that regime type is symmetric. In this section, we relax these assumptions to observe their effects on risk.

6.1. Multiple tests and overall risk

First, we analyze a case in which both states contribute to overall safety. Real-world qualitative races often proceed as a series of steps. During the Space Race, for example, the U.S. conducted a failed satellite test in December 1957 before succeeding two months later. If each test carries some risk, then the choices of *both* the winner and loser of the technology competition determine the overall level of risk. This generalization raises important theoretical questions: 1) How does the likely winner's safety investment change now that her optimal strategy depends on the likely loser's safety investments? and 2) How does the structure of the safety provision burden affect states' incentive to win the race and the overall level of risk? We explore these questions in each of the three information contexts of our baseline specification. Now, let the winner contribute a fraction $\gamma \in [0.5, 1]$ to overall safety, while the loser contributes $(1 - \gamma)$.[23] In our base model, $\gamma = 1$ and only the winner conducts a risky test of the technology. However, as $\gamma \to 0.5$, the eventual loser of the race approaches the leader's level of contribution to safety. In each state's utility function, then, the winner's expected payoff is multiplied by $\hat{s}_i$, where

$$\hat{s}_i = \gamma s_i + (1 - \gamma) s_j$$

As before, we solve the model in all three information scenarios and present the results as propositions. In the no information case, the symmetric equilibrium dictates that both states exert the same level of safety efforts as in the main section when $\gamma = 1$. When the competition loser's safety efforts–or lack thereof–also affects disaster risk, we find that the equilibrium safety efforts unambiguously decreases.

Proposition 6. *In the no information case, the unique equilibrium level of safety efforts in a symmetric BNE of pure strategies is given by:*

$$s^*_{\varnothing} = \min\left\{1, \frac{\mu[\gamma + (1-\gamma)(1-\eta)]}{2\eta[F(\mu) - F(0)]}\right\}$$

The expected level of disaster risk is then

$$D_{\varnothing} = \max\left\{0, 1 - \frac{\mu[\gamma + (1-\gamma)(1-\eta)]}{2\eta[F(\mu) - F(0)]}\right\}$$

The less safety is determined exclusively by the winner, the more the states are willing to sacrifice safety to increase the chance of winning the race. This is due to a *public-good* effect. As the fraction of the benefits of a given state's safety decrease relative to the cost, it is less willing to invest in safety over performance. We see that the amount by which the safety efforts decrease is linear in γ.

Turning to the private information scenario, states are aware of their endowed capability and how rare it is compared to the general population, while still unaware of the opponent's capabilities. In this case, the states condition safety investment strategies on their capabilities. First, we establish the equivalent of Lemma 1 for $\gamma < 1$.

Lemma 2. *In the private information scenario when $\gamma < 1$, k_i is always strictly increasing in x_i.*

Then Proposition 7 characterizes the equilibrium safety investments and disaster risk.[24]

Proposition 7. *In the private information case, the equilibrium level of safety efforts in a symmetric BNE of pure strategies is unique up to a set $\underline{x}$ and is given by:*

$$s^*_{private}(x_i) = \min\left\{1, \frac{\int_{\underline{x}}^{x_i} \Omega(c)^{\frac{\eta}{\gamma-(1-\eta)(1-\gamma)}}\, dc}{\Omega(x_i)^{\frac{\eta}{\gamma-(1-\eta)(1-\gamma)}}}\right\}$$

$$\textit{where } \Omega(x_i) = (1-\eta)(1-\gamma) + [\gamma - (1-\eta)(1-\gamma)] \cdot F(x_i)$$

The disaster risk is given by:

$$D_{private} = 1 - \frac{2}{\mu} \cdot \int_0^{\mu} \min\left\{1, \frac{\int_{\underline{x}}^{x_i} \Omega(c)^{\frac{\eta}{\gamma-(1-\eta)(1-\gamma)}}\, dc}{\Omega(x_i)^{\frac{\eta}{\gamma-(1-\eta)(1-\gamma)}}}\right\} \cdot [\gamma F(x) + (1-\gamma)(1-F(x))]dx$$

Although the expression is too complicated to elicit general intuition from, we can get insight into the effect of γ on the equilibrium effort when enmity levels are extreme. As $\eta \to 0$, and the states do not mind whether they or their opponent successfully builds the technology, Proposition 7 dictates that $s^*_{private} = 1$. When $\eta = 1$, and the states are indifferent between disaster and victory for their adversaries, $s^*_{private} = \frac{\int F(c)^{\frac{1}{\gamma}} dc}{F(x)^{\frac{1}{\gamma}}}$. Remember from Proposition 2 that when $e = 1$, the equilibrium safety effort level was $\frac{\int F(c)dc}{F(x)}$. As γ falls, more equilibrium safety comes from the loser. This produces a *public-good effect*, as γ is the fraction of a state's safety provision that it internalizes, and $1 - \gamma$ is the fraction that goes to the other state if it wins. As γ decreases, a state internalizes less of the benefit of its own safety level, causing it to reduce provision.[25] In addition, when enmity starts to tend away from 1, low capability players put more effort into safety, since even the loser's safety choice matters in her utility function.[26] This produces an additional *selection effect*; that is, compared to the case when $\gamma = 1$, moderately high capability players are now the most risky. Since they are more likely to win, this increases overall disaster risk. In sum, in an environment where the enmity level is high, having safety risk be dispersed between the winner and the loser makes states put less effort into safety.

Due to the complexity of states' utility functions in the public information case, we instead report the results of numerical simulations of equilibria to compare to the no information and private information cases. Figure 4 presents the disaster risk for $\gamma \in [0.5, 1]$ for the same parameter values we used in simulations in Figure 2 ($\eta = 0.9, \mu = 1.44$) under high ($m = 10$) and moderate ($m = 5$) values of decisiveness. We see that, as in the no information and private information cases, disaster risk is monotonically decreasing in γ in the public information case as well. Similar to the other cases, the same public-good and selection effects apply. As γ is lowered, the loser contributes more to overall risk. Since the loser is more likely to have lower capabilities than the winner, and thus also to invest less in safety in order to win, this serves to increase risk. Likewise, as other players contribute less to overall safety, each faces a temptation to shirk in their safety investments. Reduced investments in safety by others lowers the expected return on safety, even for actors that are likely to win the race.

6.2 Risk and the marginal value of winning

Up until now, we have held fixed the η parameter, which we have termed enmity and characterizes the opportunity cost of one's rival winning the race. Enmity can be high for one of two reasons. First, enmity can be high because states are existing rivals.[27] When states are more intense rivals, they are more willing to cut corners to develop a new technology. This is one of the reasons the United States government ignored some of the concerns of the director of the Los Alamos Laboratory, J. Robert Oppenheimer, over the development of the atomic bomb: they feared Germany winning the race.[28] Presumably, the U.S. would not have taken on the same level of risk had the U.K. been developing a nuclear bomb instead. Second, enmity can be high because losing the race may quite harmful to one's security if the increase in capability enabled by the technology is relatively large. In contrast, states developing a next-generation fighter jet or tank are unlikely to prompt high levels of corner cutting on safety, even competing against a bitter rival, as the expected value of losing the race is only marginally smaller than that of winning. In this case, we expect the e parameter to be relatively low. Our model captures both of these intuitions. In all three information scenarios, states choose weakly lower safety levels when the level of enmity between them is higher. Results are presented for the public information case in Figure 5 for high and low values of decisiveness m and of γ. We see that higher values of m and lower values of γ increase the concavity of the curve. That is, in highly decisive races or races in which both the winner and loser share approximately equal safety burdens, a race can quickly become maximally dangerous even when enmity is still low. In addition, we see an interaction effect: for low γ and high m, enmity is more harmful than if only one of these conditions are met.[29]

6.3 Regime type and risk

Another parameter we have kept fixed is our value of disaster, which recall is normalized to 0. However, it is likely that the relative value of a disaster depends on a state's regime type. In military conflict, for example, democracies tend to exhibit greater care to reduce casualties than autocracies (Gartzke [2001]). Likewise, in technology races, we might expect democracies to value minimizing the chance of a disaster more highly than autocratic leaders. To formalize this intuition, let a disaster reduce the value of the race outcome relative to winning by a factor of d_i, which is drawn to a distribution $d_i \sim_{iid} H(d)$ that is independent of a state's level of capabilities. Therefore, we can view democratic states as having lower values of d_i.[30] As expected, we find that if races are more likely among autocracies, as parameterized by a higher expected disaster value, risk increases in all information scenarios. Importantly, this is driven not only by races between autocracies but also by races between democracies and autocracies: in response to an autocratic rival's corner cutting on safety, a democracy will choose a lower safety level than it would when racing against another democracy. The following proposition formalizes this result:

Proposition 8. *In all information scenarios, the expected safety of the race weakly increases as the average cost of a disaster rises.*

7. Conclusion

In the introduction, we suggested that qualitative technology races present novel sources of risk that affect states' national security. Given the rate of progress on a number of powerful, risky military technologies such as advanced artificial intelligence [Brundage et al., 2018, Dafoe, 2017] and enhanced biological agents [Mukunda et al., 2009, Stern, 2002], such risks will become increasingly important factors in states' decision-making, even as they remain understudied in the international relations literature. To that end, we develop a model that seeks to understand the strategic forces influencing such risk. We find that the level of risk depends on states' knowledge about each others' capabilities as well as the decisiveness of the race. When races are less decisive, as is likely the case in the early stages of research or in novel fields, public knowledge is beneficial, preventing weak laggards from cutting corners on safety. On the other hand, at high levels of decisiveness, when the race is in its final stages or research is in a well-established field, private knowledge is safer, preventing a race to the bottom. Finally, we show that as the eventual loser is allowed to conduct more powerful tests or as enmity between players rises, overall safety falls.

A clearer understanding of the strategic forces influencing states' technology development can inform policymakers seeking to reduce the risk of an accident. First, our model implies that disclosure policies are important for novel technologies. As we have seen, public revelation of capabilities can produce a dangerous competition in which states steeply cut corners on safety to win. On the other hand, keeping capabilities private can induce risky behavior by states who believe they are lagging behind, as the U.S. did during the ICBM race. In this case, if the Soviet Union had credibly shared its capabilities, it may have prevented a dangerous race. Taken together, these two imply that there is room for policies, such as agreements to share technology or provide transfers to identify incapable laggards and prevent them from racing [Stafford and Trager, 2022]. Second, policymakers should be clear-eyed about to what extent cutting corners on safety will produce additional gains in the race. As we have seen, public knowledge of capabilities and high levels of safety are optimal when decisiveness is low. Thus, giving in to the hype about rapid, certain progress of a new technology may unnecessarily increase risk if such hype is not well-founded [Smith, 2020]. Rapid improvement in machine language-processing via so-called large-language models, for example, has produced media hype [Bender and Koller, 2020] about the imminent arrival of human-level or otherwise transformative artificial intelligence, despite experts still believing that such progress remains decades away [Zhang et al., 2022].

We believe that the findings of our model can admit a number of extensions and therefore suggest directions for future research. First, we assume that the information partitions are given exogenously. In many scenarios, however, states may choose to share information or close off development to increase their own chances of winning the race. Just as in conventional arms races, states can signal their military superiority with a public display of weapons capability in an effort to deter would-be attackers. On the flip side, states may have an incentive to engage in espionage to uncover the capabilities of their rivals or gain access to their knowledge base. Soviet spying during the race for the atomic bomb likely accelerated their program by a year or two [Ord, 2022]. Studying a model in states are allowed to disclose information voluntarily or spy on rivals will help elucidate which information scenarios are most likely. Second, our model makes two restrictive assumptions about safety research: that safety research is linear in reducing performance and that the effects of safety research are known. Trager et al. [2021] allows the safety-performance tradeoff to vary, finding that states choose higher safety levels when the tradeoff is more concave. Lower returns to safety relative to performance could also be a result of uncertainty. For a risk-averse state, more uncertainty over safety research will reduce investment in safety, making the race more dangerous. Third, our model considers decisiveness to be exogenous. Instead, we might expect it to vary over time; as we have shown, the race for the atomic bomb became more decisive–and thus far more risky–as research progressed. Therefore, extending the model to a dynamic game in which decisiveness varies over the course of the race might yield insights about when races are most risky. Fourth, it may be important to analyze the effect of information in a dynamic context where agreements are possible. Here, information plays a different role, sometimes allowing states to increase general welfare by conditioning their strategies on each other's behavior [Stafford and Trager, 2022]. Fifth, to focus attention on states' decisions to allocate their R&D budget on capabilities and safety, we have assumed budgets and therefore entry into the race are exogenous. Instead, we might consider a model in which states face a cost function and choose how much to spend on capabilities and safety, in which the optimal solution may be to simply stay out of the race. After all, the majority of states have never invested in a nuclear weapons program, and such a model would help elucadate the reasons behind this empirical regularity.

In this paper, we have sought to tie together various strands of literature–on qualitative races and on risks from various technologies such as nuclear weapons and bioweapons–into a unified model. In doing so, we hope to provide a framework upon which scholars can build as they seek to understand the strategic effects of and risks from emerging technologies. In particular, our model contributes to understanding the role of information and uncertainty in qualitative arms races showing that the decisiveness of the race can change the qualitative and quantitative effects of information. We hope that these insights are not only theoretically insightful but can be used to improve policy decisions so that advanced technologies are developed for the benefit of all.

Notes

[1] See Huntington [1958] for a related distinction between qualitative and quantitative arms races.

[2] Ellsberg [2017] notes that the Manhattan Project continued to take on unnecessary risks even after it became apparent that Germany would lose the war.

[3] Both problems are evident in current AI systems. Because are often trained on datasets that reflect human biases, current AI models often produce racist or otherwise biased outputs by default [Weidinger et al., 2021]. Likewise, the increasing cost of training state-of-the-art systems serves to "de-democratize" AI progress by shifting frontier research from universities to technology firms [Ahmed and Wahed, 2020].

[4] For a comprehensive review of the literature on arms races, see Glaser [2000].

[5] A related literature studies uncertainty over the utility functions over the value of prizes in arms races. Relevant papers include Jervis [1976], Kydd [1997], and Fearon [2011].

[6] That is, a costly peace [Powell, 1993].

[7] Note that early studies of qualitative races failed to fully appreciate this point. Focusing on relatively small innovations, Huntington [1958] argues that the main tradeoff faced by states is whether the development of qualitatively new weapons is worth a more rapid depreciation of current weapon stocks, while Intriligator and Brito [1984] find that qualitative races marginally increase the probability of war initiation relative to quantitative races.

[8] However, it must be noted that even such a draconian safety measure might not fully eliminate risk. A recent study of BSL-3 labs (the highest level of security in the U.S.) found a risk of about 1 accidental infection of a dangerous pathogen for every 100 full-time person-years of work [Lipsitch and Inglesby, 2014].

[9] We consider this a plausible simplifying assumption. Though states are able to increase the share of expenditures devoted to a particular technology, qualitative races are generally brief enough that they cannot greatly increase their military budget in the same time period [Ord, 2022].

[10] Though we conceive of the benefits of a new technology as economic or military, qualitative races may also accrue prestige benefits to the winner [Barnhart, 2021].

[11] Certainly, igniting the atmosphere would be equally bad for all states!

[12] In contrast, Bas and Coe [2016] model research success as independent of the level of R&D effort.

[13] Though Conrad and Spaniel [2021] note that a logistic CSF does not display decreasing returns to scale, the importance of winning a qualitative race to a state is likely to mean that in principle, such a constraint will not strongly influence state behavior. Consider that from 1964 to 1966, the U.S. was willing to spend over 4% of its federal budget on NASA to win the space race against the Soviet Union.

[14] Note that this assumption allows us to nest our model as a more general case of the model in Armstrong et al. [2016], since contests of this form converge uniformly to all-pay auctions [Jia et al., 2013, Che and Gale, 2000].

[15] In models of conflict, m represents the offensive advantage [Fearon, 2018].

[16] Proofs of all propositions are presented in Appendix A.

[17] This is a common technique for finding closed-form expressions for equilibria in difference-form contests [Ryvkin and Drugov, 2020], analogous to using Gumbel-distributed noise to characterize choice probabilities in logit regressions [McFadden, 1974].

[18] For the public information case, we a modified version of the computational solution implemented by Muller et al. [2021].

[19] Though our statement for public information is relatively weaker, we note that in simulations of "reasonable" parameter values, such as those simulated in Figure 2. risk sharply increases with decisiveness in the public information case as well, only mildly decreasing as m increases without bound above 20.

[20] Note that a larger μ implies that relative capability is likely more significant than relative corner-cutting in determining the winner of the contest.

[21] In Appendix B, we plot each gap in risk between information states.

[22] Though the Space Race likely entailed a lower overall level of risk than many other technology races, tragedies such as the *Challenger* disaster show that technology races for space exploration are not without risk.

[23] This is equivalent to the weighted sum provision function in public goods problems [Buchholz and Sandler, 2021].

[24] Unlike in Proposition 2, $s^*_{private}(x_i)$ is not an equilibrium for all $x_i \in supp(C)$. In Appendix A, we establish the uniqueness of $\underline{x}$ up to (m, η, μ, γ) and the behavior of $x_i < \underline{x}$.

[25] In the context of climate policy, Sandler [1998] notes that, states have achieved a far higher reduction in sulfur emissions than in nitrogen oxide emissions since 1985 due to the fact that γ is far higher in the former case.

[26] Indeed, for low enough η, γ, low capability players deviate to the corner solution $s^*(x) = 1$. For example, for $\gamma < \frac{1-2\eta}{2-\eta}$, all $x_i \leqslant \mathbb{E}[x_i] = \frac{\mu}{2}$ play $s^*(x) = 1$.

[27] This is consistent with common approaches to rivalry in international relations theory. See, for instance, Hensel et al. [2000] and Goertz and Diehl [1995].

[28] Ord [2020].

[29] Results for the no information and private information cases are similar and are presented in Appendix B.

[30] Although there is a positive correlation between income and democracy (see e.g. Treisman [2020]), this may not hold conditional on states being in a technology race.

Figures and Tables

		Gap with Capable Leader	
		Small	Large
Information Scenario	Private	Low (S. Africa nuclear program)	High (ICBM race)
	Public	High (Space Race)	Low (Nuclear NPT)

Table 1: Risk in historical technology races

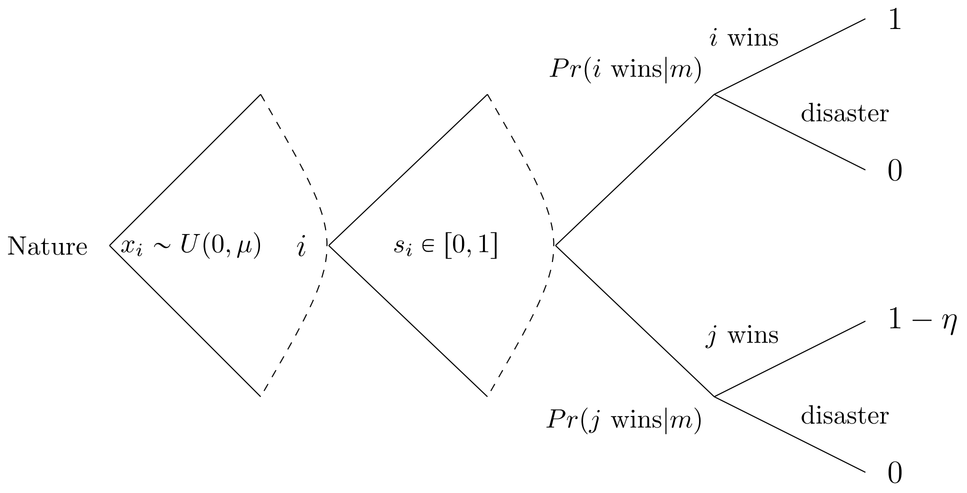

Figure 1: The game tree

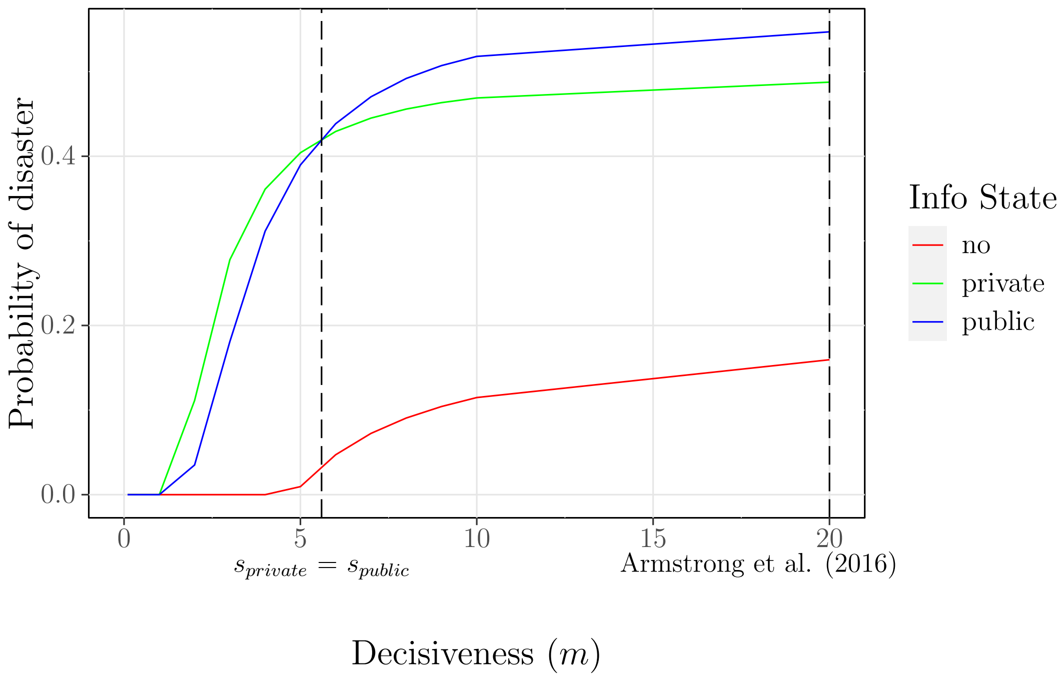

Figure 2: Disaster risk ($\mu = 1.44, \eta = 0.9$)

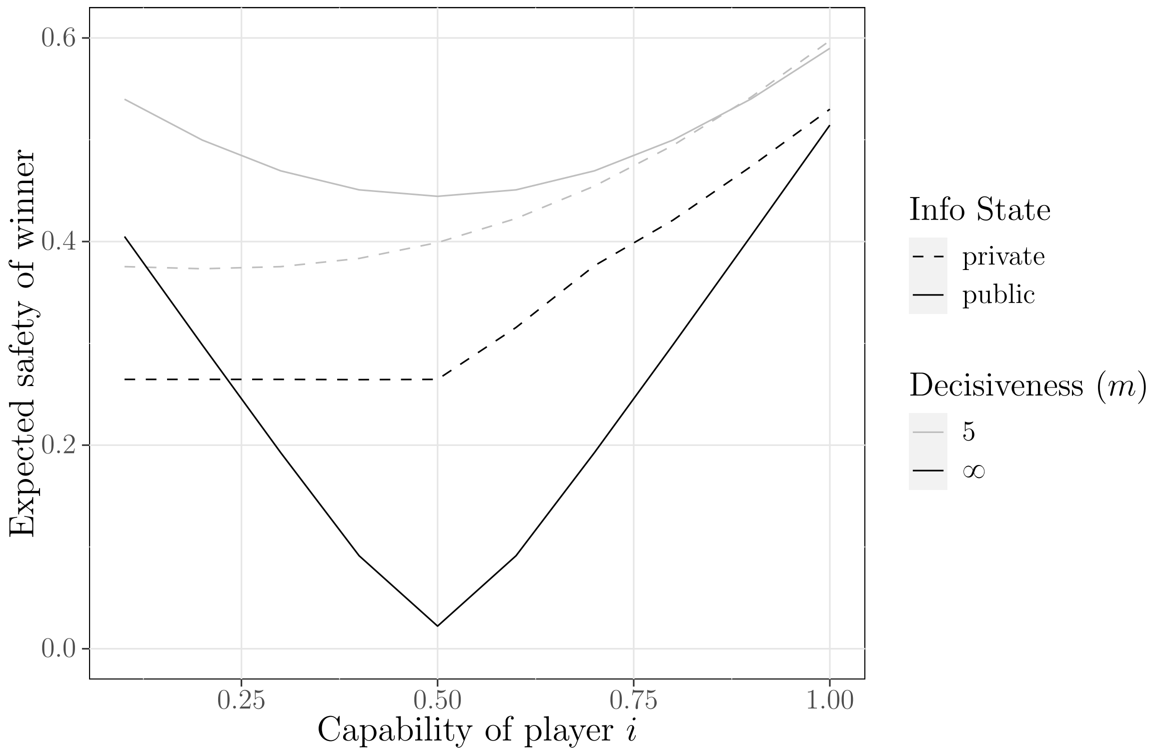

Figure 3: Expected safety of the race winner ($\mu = 1, \eta = 1, x_j = 0.5$)

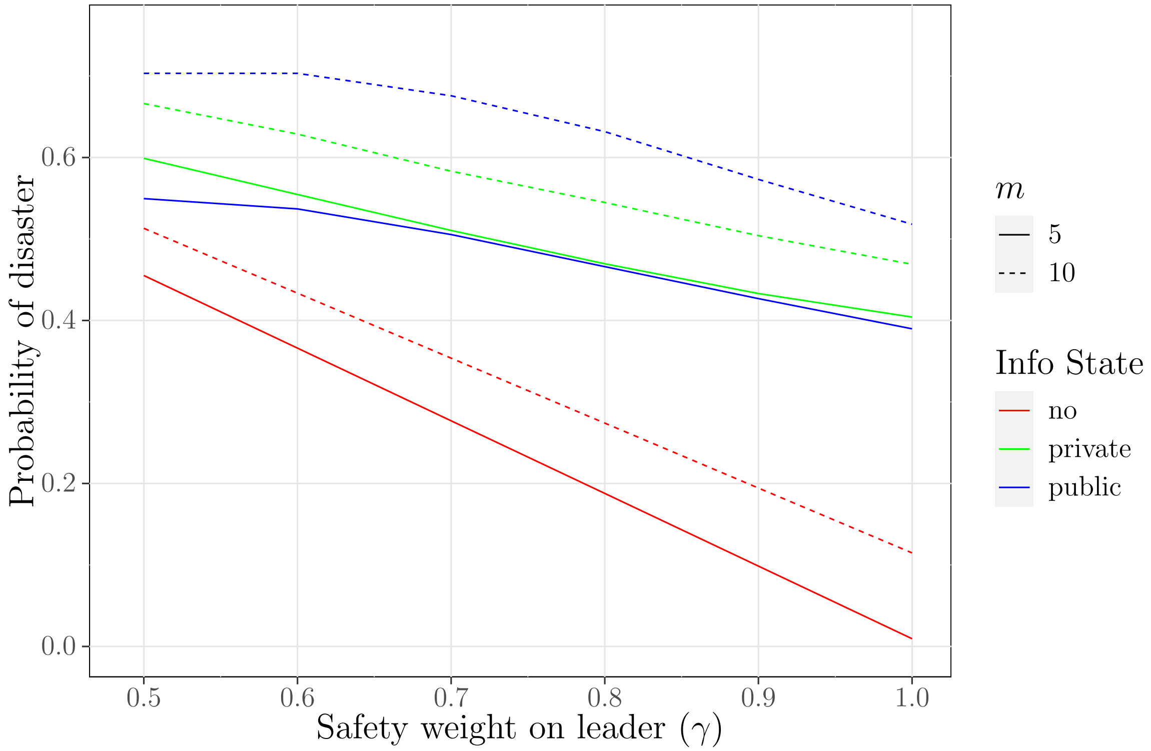

Figure 4: Varying safety contributions of the winner

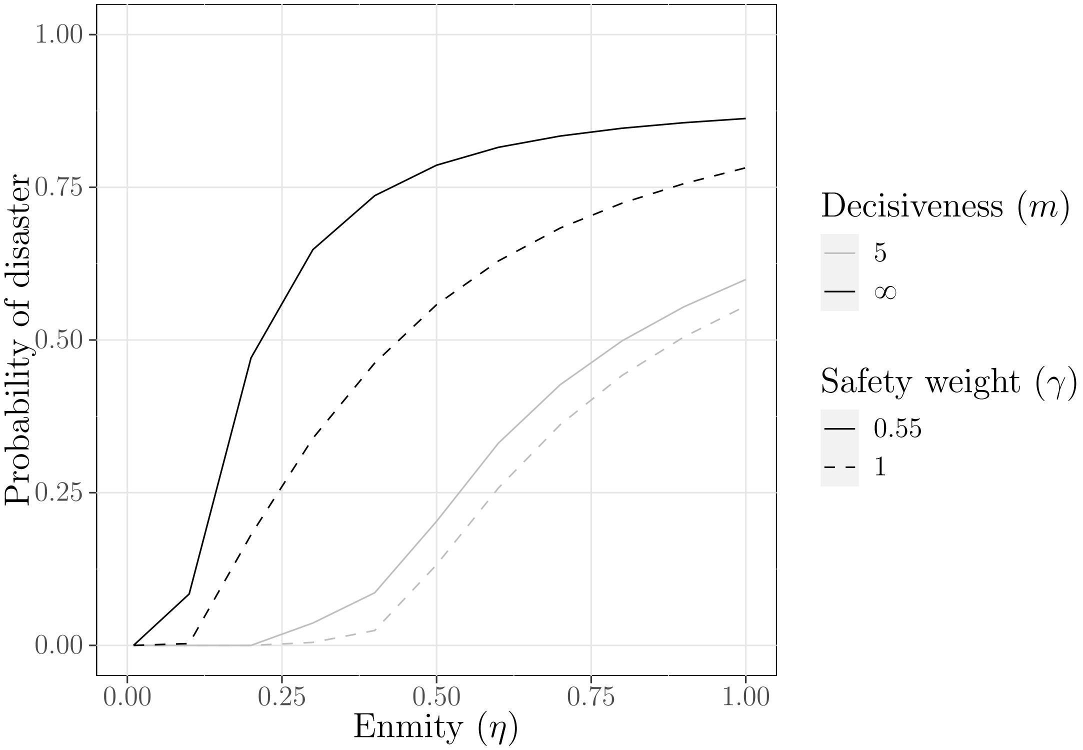

Figure 5: Effects of enmity under public information ($\mu = 0.72$)

References

Nur Ahmed and Muntasir Wahed. The de-democratization of ai: Deep learning and the compute divide in artificial intelligence research. *arXiv preprint arXiv:2010.15581*, 2020.

AP. Putin: Leader in artificial intelligence will rule world, September 2017. URL `https://www.cnbc.com/2017/09/04/putin-leader-in-artificial-intelligence-will-rule-world.html`. Section: Technology.

Stuart Armstrong, Nick Bostrom, and Carl Shulman. Racing to the precipice: a model of artificial intelligence development. *AI & Society*, 31(2):201–206, May 2016.

Joslyn Barnhart. The Dynamics of Prestige Races: What the Space Race Means for the Future of Technological Development. 2021.

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

Michael R. Baye and Heidrun C. Hoppe. The strategic equivalence of rent-seeking, innovation, and patent-race games. *Games and Economic Behavior*, 44(2):217–226, August 2003.

Emily M Bender and Alexander Koller. Climbing towards nlu: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 5185–5198, 2020.

Martin Beraja, Andrew Kao, David Y Yang, and Noam Yuchtman. Ai-tocracy. *The Quarterly Journal of Economics*, 138(3):1349–1402, 2023.

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

Miles Brundage, Shahar Avin, Jack Clark, Helen Toner, Peter Eckersley, Ben Garfinkel, Allan Dafoe, Paul Scharre, Thomas Zeitzoff, Bobby Filar, et al. The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. *arXiv preprint arXiv:1802.07228*, 2018.

Wolfgang Buchholz and Todd Sandler. Global Public Goods: A Survey. *Journal of Economic Literature*, 2021.

Stephen Cave and Seán S ÓhÉigeartaigh. An ai race for strategic advantage: rhetoric and risks. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 36–40, 2018.

Yeon-Koo Che and Ian Gale. Difference-Form Contests and the Robustness of All-Pay Auctions. *Games and Economic Behavior*, 30(1):22–43, January 2000.

Justin Conrad and William Spaniel. *Militant Competition: How Terrorists and Insurgents Advertise with Violence and How They Can Be Stopped.* Cambridge University Press, September 2021. ISBN 978-1-108-99828-4.

Allan Dafoe. AI Governance: A Research Agenda. *Future of Humanity Institute*, July 2017. URL `https://www.fhi.ox.ac.uk/wp-content/uploads/GovAIAgenda.pdf`.

Daniel Ellsberg. *The Doomsday Machine: Confessions of a Nuclear War Planner*. Bloomsbury Publishing USA, December 2017. ISBN 978-1-60819-670-8.

James D Fearon. Rationalist explanations for war. *International Organization*, 49(3):379–414, 1995.

James D Fearon. Arming and arms races. *Annual Meetings of the American Political Science Association, Washington, DC*, 2011.

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

Mark Fey and Kristopher W. Ramsay. Uncertainty and incentives in crisis bargaining: Game-free analysis of international conflict. *American Journal of Political Science*, 55(1): 149–169, 2011.

Matthew Fuhrmann and Yonatan Lupu. Do Arms Control Treaties Work? Assessing the Effectiveness of the Nuclear Nonproliferation Treaty. *International Studies Quarterly*, 60 (3):530–539, September 2016. Publisher: Oxford Academic.

Erik Gartzke. Democracy and the preparation for war: Does regime type affect states' anticipation of casualties? *International Studies Quarterly*, 45(3):467–484, 2001.

Charles L Glaser. The causes and consequences of arms races. *Annual Review of Political Science*, 3(1):251–276, 2000.

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

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

Jack Hirshleifer. Chapter 7 Theorizing about conflict. In *Handbook of Defense Economics*, volume 1, pages 165–189. Elsevier, January 1995.

Michael Dean Horn. *Arms races and the international system.* Ph.D., University of Rochester, United States – New York, 1987.

Michael C Horowitz. Artificial Intelligence, International Competition, and the Balance of Power. *Texas National Security Review*, 1:22, 2018.

Samuel P Huntington. Arms races-prerequisites and results. *Public Policy*, 8:41–86, 1958.

Nuclear Threat Initiative. United States Missile Overview, 2022. URL `https://www.nti.org/analysis/articles/united-states-missile/`.

Michael D. Intriligator and Dagobert L. Brito. Can Arms Races Lead to the Outbreak of War? *Journal of Conflict Resolution*, 28(1):63–84, March 1984. Publisher: SAGE Publications Inc.

Robert Jervis. *Perception and Misperception in International Politics.* Princeton University Press, Princeton, 1976. Book.

Hao Jia, Stergios Skaperdas, and Samarth Vaidya. Contest functions: Theoretical foundations and issues in estimation. *International Journal of Industrial Organization*, 31(3): 211–222, May 2013.

Andrew Kydd. Sheep in sheep's clothing: Why security seekers do not fight each other. *Security Studies*, 7(1):114–155, 1997.

Andrew Kydd. Trust, reassurance and cooperation. *International Organization*, 54(2):325–57, 2000.

Andrew H Kydd and Scott Straus. The road to hell? third-party intervention to prevent atrocities. *American Journal of Political Science*, 57(3):673–684, 2013.

Daniel J. Lawler and Erin R. Mahan. Foreign Relations of the United States, 1961â1963, Volume VIII, National Security Policy, June 1961. URL `https://history.state.gov/historicaldocuments/frus1961-63v08/d29`.

Peter Liberman. The rise and fall of the south african bomb. *International Security*, 26(2): 45–86, 2001.

Marc Lipsitch and Thomas V. Inglesby. Moratorium on Research Intended To Create Novel Potential Pandemic Pathogens. *mBio*, 5(6):e02366–14, December 2014. Publisher: American Society for Microbiology.

Jennifer G Mathers. 'a fly in outer space': Soviet ballistic missile defence during the khrushchev period. *The Journal of Strategic Studies*, 21(2):31–59, 1998.

Daniel McFadden. The measurement of urban travel demand. *Journal of Public Economics*, 3(4):303–328, November 1974.

Adam Meirowitz and Anne Sartori. Strategic Uncertainty as a Cause of War. *Quarterly Journal of Political Science*, 3(4):327–352, December 2008.

Gautam Mukunda, Kenneth A Oye, and Scott C Mohr. What rough beast? synthetic biology, uncertainty, and the future of biosecurity. *Politics and the Life Sciences*, 28(2): 2–26, 2009.

Jonas Muller, Paolo Bova, Ben Harack, Tanja Ruegg, Jasmine Brazilek, Vasily Kuznetsov, and Miles Tidmarsh. Baseline web app, June 2021. URL `https://www.modelingcooperation.com/software`.

Vipin Narang. Strategies of Nuclear Proliferation: How States Pursue the Bomb. *International Security*, 41(3):110–150, January 2017.

Wim Naude and Nicola Dimitri. The race for an artificial general intelligence: implications for public policy. *AI & Society*, 35(2):367–379, June 2020.

Newsweek. Einstein, the Man Who Started it All. *Newsweek*, March 1947.

Toby Ord. *The Precipice: Existential Risk and the Future of Humanity*. Hachette Books, March 2020. ISBN 978-0-316-48489-3.

Toby Ord. Lessons from the development of the atomic bomb. Technical report, Centre for the Governance of AI, 2022.

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

Robert Powell. Bargaining and learning while fighting. *American Journal of Political Science*, 48(2):344–361, 2004.

Or Rabinowitz and Nicholas L Miller. Keeping the bombs in the basement: Us nonproliferation policy toward israel, south africa, and pakistan. *International Security*, 40(1):47–86, 2015.

Kristopher W. Ramsay. Information, Uncertainty, and War. *Annual Review of Political Science*, 20(1):505–527, May 2017.

William Reed. Information, Power, and War. *American Political Science Review*, 97(4): 633–641, November 2003.

Lewis F Richardson. *Arms and insecurity: A mathematical study of the causes and origins of war*. Boxwood Press, 1960.

Michelle Rozo and Gigi Kwik Gronvall. The Reemergent 1977 H1N1 Strain and the Gain-of-Function Debate. *mBio*, 6(4):e01013–15, August 2015. Publisher: American Society for Microbiology.

Stuart Russell. *Human Compatible*. Penguine Books, 2019.

Dmitry Ryvkin and Mikhail Drugov. The shape of luck and competition in winner-take-all tournaments. *Theoretical Economics*, 15(4):1587–1626, 2020.

Susan G. Sample. Arms Races and Dispute Escalation: Resolving the Debate. *Journal of Peace Research*, 34(1):7–22, February 1997. Publisher: SAGE Publications Ltd.

Todd Sandler. Global and Regional Public Goods: A Prognosis for Collective Action. *Fiscal Studies*, 19(3):221–247, 1998.

Thomas C Schelling. *The Strategy of Conflict*. Harvard University Press, Cambridge, MA, 1980.

Stergios Skaperdas. On the formation of alliances in conflict and contests. *Public Choice*, 96 (1):25–42, July 1998.

Frank L Smith. Quantum technology hype and national security. *Security Dialogue*, 2020.

Eoghan Stafford and Robert Trager. The iaea solution for risky technology races: Knowledge sharing to reduce competition and proliferation. *Working Paper*, 2022.

Eoghan Stafford, Robert Trager, and Allan Dafoe. International Strategic Dynamics of Technology Races With Safety-Performance Tradeoffs. *Working Paper*, 2021.

Jessica Stern. Dreaded Risks and the Control of Biological Weapons. *International Security*, 27(3):89–123, 2002. Publisher: The MIT Press.

Robert Trager, Paolo Bova, Eoghan Stafford, Allan Dafoe, and Nicholas Emery-Xu. Welfare implications of safety-performance tradeoffs in ai safety research, February 2021.

Daniel Treisman. Economic development and democracy: predispositions and triggers. *Annual Review of Political Science*, 23(1):241–257, 2020.

Michael D. Wallace. Arms Races and Escalation: Some New Evidence. *Journal of Conflict Resolution*, 23(1):3–16, March 1979. Publisher: SAGE Publications Inc.

Michael D. Wallace. Armaments and Escalation: Two Competing Hypotheses. *International Studies Quarterly*, 26(1):37–56, March 1982.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*, 2021.

Martin L. Weitzman. Recombinant Growth. *The Quarterly Journal of Economics*, 113(2): 331–360, May 1998.

Donald Wittman. Bargaining in the Shadow of War: When Is a Peaceful Resolution Most Likely? *American Journal of Political Science*, 53(3):588–602, 2009.

Eliezer Yudkowsky et al. Artificial intelligence as a positive and negative factor in global risk. *Global catastrophic risks*, 1(303):184, 2008.

Baobao Zhang, Noemi Dreksler, Markus Anderljung, Lauren Kahn, Charlie Giattino, Allan Dafoe, and Michael C. Horowitz. Forecasting AI Progress: Evidence from a Survey of Machine Learning Researchers, June 2022. URL `http://arxiv.org/abs/2206.04132`. arXiv:2206.04132 [cs].