Information Hazards in Races for Advanced Artificial Intelligence

[Preliminary draft: Do not cite]

Nicholas Emery*, Andrew Park†, Robert Trager‡

June 2022

Abstract

We study how the information environment affects races to implement a powerful new technology such as advanced artificial intelligence. In particular, we analyze a model in which a potentially unsafe technology may cause a "disaster" that affects all actors and actors that implement the technology face a tradeoff between the safety of the technology and their performance in the race. Combining analytic and computational approaches, 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 information hazard result of Armstrong et al. [2016]–that the public information scenario is more dangerous than the private information scenario–only holds under high decisiveness. As decisiveness decreases, the

*UCLA, Department of Economics

†UCLA, Department of Economics

‡UCLA, Department of Political Science. We would like to thank Paolo Bova and Jonas Mueller for helpful feedback and comments. We note our reliance for some derivations on notes belonging to Modelling Cooperation. We thank the Stanford Existential Risks Initiative and the Open Philanthropy Project for financial and logistical support. All remaining errors are our own.

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 moral hazard effect.

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 such technology and arms races may become a key feature of international politics in the coming decades, as states race to be the first to develop new technologies such as advanced artificial intelligence (AI) or nanotechnology that will give them a sudden increase in capability over other states [Bostrom, 2014].

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 preempt 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. Biological weapons use has a relatively high

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

probability of infecting the user, which is 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 or even existential risks are involved in such arms races [Ord, 2020, Bostrom, 2019, Farquhar et al., 2017].[2]

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].

We examine a model that captures these dynamics. Combining analytic and computational approaches, 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

[2]Aschenbrenner [2020] argues that the rate of the development of such technologies may be increasing in the short run.

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 information hazard result of Armstrong et al. [2016]–that the public information scenario is more dangerous than the private information scenario–only holds 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 moral hazard effect.

Some of these dynamics are illustrated in the race for the first nuclear bomb during World War II when certain physicists involved in the Manhattan Project expressed concerns over the safety-performance tradeoff. Edward Teller, for example, feared that a nuclear fusion reaction would produce a temperature exceeding that of the Sun (15,000,000°C), igniting the atmosphere and ending life on Earth. He privately urged the US 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 US government's uncertainty over the level of progress of Germany's development of nuclear weapons.[3]

A number of scholars believe that the development of advanced forms of artificial intelligence (transformative AI, or TAI) will exhibit similar strategic dynamics [Bostrom, 2014,

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

Yudkowsky, 2013]. Bostrom [2017] highlights two classes of risks in such scenarios: the risk of misaligned AI objective functions (the control problem), and the use of TAI by actors wishing to impose harms on others (the political problem). Contemporary AI systems have already exhibited such pathologies [Beraja et al., 2020, Mehrabi et al., 2021]. As the rewards from advanced AI become more apparent, firm and state actors may have incentives to increase the pace and secrecy of development, with as-yet-unknown effects on risk. Currently, much AI development is governed by a norm of openness, with new research published on the open-source repository arXiv. However, that norm is changing. For example, OpenAI decided against publishing the source code for its GPT-3 language model to ensure that it could not be misused by malicious actors. Bostrom [2017] notes that competitive pressure could lead firms or states to close off development entirely, preventing proper oversight by the public or other benevolent actors.

According to Armstrong et al. [2016], a closed information environment could be seen as wholly beneficial. They study a model of a qualitative arms race in which actors face a safety-performance tradeoff, and find that under certain conditions, shared information about relative capabilities decreases welfare.[4] In their model, n competitors compete to in a technology race while aware that a disaster might be caused by the winner of the race, or the owner of the technology. Endowed with innate capability x_i, each competitor i can invest in safety (s_i), which negatively affects their performance level ($x_i - s_i$). The competitor j with the highest competence level wins the race and builds the technology, which succeeds with probability s_j and results in a disaster with probability $1 - s_j$. Their most surprising

[4]Bostrom [2011] defines an information hazard as a piece of information whose dissemination increases the expected harm to society.

result is that, counter-intuitively, when the actors are sufficiently rivalrous, the risk level of the race rises as actors learn more about their own capability and that of their opponents. Although the result holds for a large fraction of parameter values in their model, we develop a model to show that this finding depends on a few extreme assumptions: 1) that the winner is absolutely determined by her performance level ($x_i - s_i$) and 2) that the winner is the sole source of risk.

To illustrate the extremity of the first assumption, note that in the Armstrong et al. [2016] model, if one actor has even ϵ greater level of performance compared to the other, she wins the race with certainty. Moreover, this infinitesimal advantage in performance becomes common knowledge in the public information case, and the fully informed players can therefore disproportionately reduce investment in safety. We instead argue that winning real-world technology races depends on a variety of factors rather than any one well-known measure of performance. For example, in the TAI case, progress may be achieved by increasing computing power via scaling laws [Hernandez, 2018], or via different approaches to algorithmic improvement. Performance is thus a function of capability (defined as a function of current knowledge, effort and resources), but also of resource allocation decisions whose effect on success is probabilistic from the point of view of the race participants. Technology development processes are inherently stochasitic in terms of the timing of advances and with respect to the victor in the competition. Victory is influenced by a variety of decisions, but in expectation not generally fully determined by any one. Thus, in contrast to the Armstrong et al. [2016] model, victory in the competition is stochastic even if one actor has slightly greater human and material resources than others or invests slightly less in the safety of the technology.

To formalize this stochasticity, we utilize a difference-form contest success function parametrized by a decisiveness parameter m. We find that the information hazard–the finding that public information is a more dangerous setting than private information–only holds under extreme values of m (indeed, the model in Armstrong et al. [2016] corresponds to the case where $m = \infty$). We find that the higher the decisiveness of the race, the more likely players are to cut corners on safety. This makes intuitive sense, as decisiveness controls the expected returns to investing in performance. Thus, if the path to AI is noisy, races will be relatively less dangerous than the highly-decisive race modelled in Armstrong et al. [2016].

We show that, under more moderate levels of the decisiveness parameter m, the public information case is often safer than the private information case. This is a result of the changing balance of two opposing forces facing states as the race becomes more noisy. The first force is that laggards are more likely to win as decisiveness decreases, which increases the overall risk because laggards take on more risk. The other is that some actors react to this lower probability of winning by increasing safety investments, as corner-cutting generates lower expected returns. As decisiveness decreases, in the public information case, the latter effect dominates for all players except those who are close in capability. Thus, across most of the type distribution, as decisiveness decreases, players increase safety investments substantially. On the other hand, in the private information case, the former effect dominates. In this case, since players are uncertain of where they are in the type distribution, they are still willing to take on risk even as decisiveness falls. Overall, then, even in cases when the public information case is the most dangerous in highly decisive races, as the returns to investing in relative capability become noisier, this dynamic may be reversed; the private information case will generally be the most dangerous.

The second assumption is likewise extreme. In the Armstrong et al. [2016] model, the sole source of risk is the winner's lack of investment in safety. This is unlikely to be the case, as even the losers' unaligned AI systems have the potential for significant harm. Generalizing beyond AI, technology races may produce other externalities as well, for example by increasing carbon emissions [Wiblin]. In order to address this possibility, we introduce a weighted sum model of safety investment, in which the overall risk of disaster comes from a weighted sum of the safety investments of the winner and the laggards. We parametrize this weighted sum by a single measure γ, the contribution of the winner's safety investments to the overall risk, while the $n-1$ laggards' safety investments contribute $1-\gamma$ in total. We find that for plausible values of γ, disaster risk is monotonically increasing as more weight is placed on the laggard. This is due to a combination of two forces. First is a *selection effect*. Having more of the overall safety burden fall on the losers increases risk because losers are likely to be laggards with relatively low capability who cut corners on safety more than capable leaders. Second is a *moral hazard* effect. If overall safety depends on a sum of each player's individual efforts, players have an incentive to free ride off the safety investments of others, a classic problem in the public goods literature [Buchholz and Sandler, 2021]. That is, if AI safety provision is more like global carbon emissions reduction than like building a dyke, we should expect overall safety provision to be lower.

Finally, we analyze the robustness of our results to other parameters in our model: enmity and the distribution of performance. We find that the results in Armstrong et al. [2016] about these parameter values still qualitatively hold but may be stronger depending on decisiveness and the safety contributions of the laggard. In all information states, the lower the variance of the performance type distribution and the higher the level of enmity, or rivalry between

the players, the riskier the race becomes. Further, we demonstrate that the effect of enmity exacerbated when decisiveness is high and the safety weight of the leader is low. That is, the effects of bitter rivalry are worse when actors know that cutting corners on safety would more likely lead them to win the race or when they do not internalize the full benefits of safety investments.

2. Information, arms racing, and risk

The sorts of incomplete information that drive the risk of conflict, climate disaster, and other public "bads," 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, drawing on the insights in behavioral economics, invokes such causes as players' 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 equilbria only obtain when the joint cost of war is large enough [Fey and Ramsay, 2011].

In contrast to the study of bargaining and war, the study of information and uncertainty

in arms races has been hindered by the lack of a standard model for thinking about such competitions.[5] Kydd [2000] and Meirowitz [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 [2008] endogenizes 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. Debs and Monteiro [2014] endogenize the choice of investment in capabilities. In contrast to Fearon [1995], they show that war is only possible when the arming state possesses private information about its level of capabilities. 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.[6]

The other literature in which our paper is situated is the economics literature on contests, in which actors race for a prize. Specifically, we focus on the literature that uses contest success functions, first studied by Tullock [2005] and later axiomatized by Skaperdas, to study the rate of returns to effort in winning the contest. In general, these models study variations of a simple utility function

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

[6]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].

$$u_i(x_i) = V_i p_i(x_i, \mathbf{x}_{-i}, m) - h_i(x_i) \tag{1}$$

in which V_i is the valuation of the prize to player i, x_i is player i's investment in the contest, which carries a cost $h_i(x_i) \geqslant 0$. p_i is the contest success function, parametrized by decisiveness parameter m to control for the returns to effort. Relatively few models in this literature, however, focus on studying uncertainty over $\mathbf{x}_{-i}$, partly due to the difficulty of finding a closed-form expression for equilibrium investment levels in such contests. Baik [1994] was the first to derive an equilibrium for the complete information case in a Tullock contest, finding that weaker players invest higher levels of effort. Grossmann [2014] studies a contest in which players' returns to effort are drawn from a Bernoulli distribution and both players share the same prior over their own and their rival's type. He finds that lower expected returns reduce both players' effort and increase expected profits. Einy et al. [2015] derive the conditions under which a contest has a pure strategy Bayesian Nash equilbrium, and Ewerhart and Quartieri [2020] derive the conditions under which a pure strategy Nash equilibrium is unique.[7]

We bring insights from these two strands of literature to study an oft-neglected type of arms competition: the qualitative arms race [Huntington, 1958]. A small but growing literature has emerged in recent years in response to two growing concerns over new military technologies. The first is that such technologies will lead to discontinuous power shifts,

[7]Generalizing from static contests, Bimpikis et al. [2019] study a dynamic contest in which players lagging in the race who lack information about the progress of their rivals exert more effort compared to players who know they are laggards. However, the static setting of our model does not permit us to analyze the role of information on such dynamics.

increasing the intensity of arms races and potential for conflict. This view is epitomized by President 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." [noa, 2017]. A second concern is that the returns to winning such contests can lead actors to cut corners in their development, potentially leading to global, even existential, risks. In the literature on qualitative races are Naude and Dimitri [2020], who show that taxing AI development and using public procurement can incentivize cooperation and reduce risk.[8] Stafford et al. [2021] analyze a dynamic arms race in which disaster 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. Unlike our model, all of these models study risk under the assumption that information about players' capabilities is common knowledge. However, we know that an additional source of risk is uncertainty over capabilities [Armstrong et al., 2016]. It is to this uncertainty that we now turn.

3. Model primitives

In our model, $n = 2$ agents (firms or states) compete to build a significant technology, which for ease of reference, we label transformative AI (TAI). Each agent i is endowed with capability level x_i, which, depending on the information state, may be unknown, privately known, or publicly known. Their capability is drawn independently from a commonly-

[8]Note that such mechanisms presume that TAI will be developed within one state. Public goods problems are exacerbated in anarchy. For a literature review, see Buchholz and Sandler [2021].

known distribution $G(x_i) = Uniform(0, \mu)$, and each player chooses safety investment level $s_i \in [0, 1]$ that determines her overall performance k_i, given by $k_i = x_i - s_i$. The player i who wins the race then implements TAI; with probability s_i, implementation is successful, and with probability $1 - s_i$, a disaster is incurred. We normalize the value of winning the race to 1 and the value of a disaster to 0. Following the standard arms race literature, we assume that each player has a symmetric level of enmity ($e \in [0, 1]$) toward her rival, which represents the opportunity cost of losing the race. Players' expected utility functions are then given by

$$u_i(s_i) = s_i Pr(i \text{ wins} | k_i, k_j) + (1 - e)s_j Pr(j \text{ wins} | k_i, k_j)$$

As noted in the introduction, the actor with the highest level of performance wins the race, and which player has the highest level of performance is partly stochastic and partly a function of the players' capabilities. Other factors, including luck or unobserved measures of capability, may also be correlated with a player's ability to win the race. To formalize these microfoundations and aid in computation of equilibrium strategies, we model each player as endowed with a known level of performance x_i plus an unknown additive noise component v_i drawn from a commonly-known distribution $V \sim Gumbel(1, \frac{1}{m})$. In addition, each player knows her rival also independently draws an additive noise component $v_j \sim V$. Therefore, we can represent each players *uncertainty-adjusted capability* with the random variable $c_i(x_i)$. This level c_i, unbeknownst to players in the game, represents the randomness of the race and determine which type wins. That is, the player with the highest value of c_i wins the

race, where $c_i(x_i) \equiv x_i + v_i - v_j$, $v_i, v_j \sim Gumbel(1, \frac{1}{m})$. We denote $F(c_i)$ and $f(c_i)$ as its distribution and density functions, respectively.[9]

We know that a contest with Gumbel-distributed noise components is equivalent to a standard Tullock [2005] contest with a logistic contest success function [Ryvkin and Drugov, 2020]. We can then rewrite the utility function as

$$\begin{aligned} u_i(s_i) &= s_i \left(\frac{\mathrm{e}^{mk_i}}{\sum_j \mathrm{e}^{mk_j}} \right) + (1-e)s_j \left(1 - \frac{\mathrm{e}^{mk_i}}{\sum_j \mathrm{e}^{mk_j}} \right) \\ &= s_i \mathbb{1}\{c_i > c_j\} + (1-e)s_j \mathbb{1}\{c_i < c_j\} \end{aligned}$$

Contest success functions have been used in both the economics literature on innovation contests [Baye and Hoppe, 2003] and the international relations literature on conflict [Skaperdas, 1998, Hirshleifer, 1995]. We choose the logistic (difference form) CSF over the ratio CSF for both tractability and theoretical reasons.[10] First, when taking expectations over CSFs, the ratio form integrates over values of 0 in the denominator such that the integral diverges. Second, [Hirshleifer, 1995] argues based on a series of historical examples, that the ratio form is only applicable to conflict when conditions are ideal; when conditions are imperfect, the difference form is more appropriate.

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. One interpretation of it is the degree of uncertainty over possible paths to TAI. In the economics of innovation

[9] A closed form derivation of the distribution is provided in Appendix A.
[10] See Skaperdas for an axiomatization.

literature, scholars have modeled idea generation as a function of existing ideas. Agrawal et al. [2019b] note that new ideas are formed as a combination of existing ideas.[11] If the search space of paths to TAI is large relative to the rate at which researchers can navigate it, then such a race is likely to have a lower decisiveness parameter. One could imagine this is the case, for example, if researchers are uncertain about which algorithms lead to TAI. However, if the search space is small relative to the rate of researchers' idea generation, the race is likely to be more decisive. This might be the case, for example, if most of the progress on machine learning benchmarks continue to come from scaling up computing power. If TAI is developed as a result of scaling up compute using known algorithms, then the current gap in capabilities between the leader and the laggard is likely to be more decisive.

Our other main focus is the weighted sum form of safety provision. Weighted sum production technologies have been studied extensively in the public goods literature.[12] Each player i chooses s_i, but the overall level of safety provision is given by $\hat{s}_i$, a weighted sum of the leader's safety provision and laggards' provision:

$$\hat{s}_i = \gamma s_i + \frac{1-\gamma}{n-1} \sum_{j \neq i} s_j \tag{2}$$

where $\gamma \in [0, 1]$. The γ parameter thus controls the proportion of risk that comes from the leader's safety investment. In some TAI development scenarios, all of the players may be developing relatively safe AI when one player discovers a new algorithm that immediately gives them TAI. In such a case, we might expect γ to be close to 1 if the actors could copy a

[11] In their model, which nests the classic Jones [1995] endogenous growth model, a researcher with access to $A^{\phi}, \phi \in (0, 1)$ ideas faces $2^{A^{\phi}}$ possible ideas combinations, only some of which are productive.

[12] See Buchholz and Sandler [2021] for a review.

safe implementation. In other cases, the increase in capabilities leading to TAI may be more gradual, such that even the laggards have relatively high capabilities when the race is over. In such a case, γ is likely to be closer to $\frac{1}{n}$.

Finally, we note that these assumptions allow us to nest the Armstrong et al. [2016] model as a special case of our more general model, allowing us to perform comparative statics on decisiveness and the safety weight of the leader. Specifically, their model is equivalent to our when $m \to \infty, \gamma = 1$. First, prior work has shown that $F(c)$ converges uniformly to $F(x)$ as $m \to \infty$ [Jia et al., 2013, Che and Gale, 2000]. That is, logistic contests converge to all-pay auctions in the limit of decisiveness, in which the player with the largest k_i wins the race with certainty. Second, note that when $\gamma = 1$, $\hat{s}_i = s_i$ in players' utility functions. We now turn to our solution concept.

4. Base model

Here we derive results for our main model under three information conditions, allowing us to perform comparative statics on the safety levels of leaders and laggards, the disaster risk, and the information hazard or increased risk of public over private information about rivals' capabilities.

4.1 No information

In this scenario, no agent knows her own capability. This scenario maps to a situation where players are unaware of how their current stock of resources maps onto the ability to make

progress toward TAI. This is distinct from the uncertainty that comes from decisiveness. With low values of decisiveness, players may know their own performance level and can channel their resources toward developing TAI, with uncertain results. Here, players also have no information about their own capability. This is a more fundamental source of uncertainty: does a player's stock of resources even contribute to AI progress at all? Realistically, then, the no information case represents a lower bound on players' knowledge, as in the real world players are likely to have at least an understanding of how to build strong AI. Because players 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 player capabilities.

Recalling that $F(c), f(c)$ represent the distribution and density of the noise-adjusted capability random variable C, we derive the unique Bayesian Nash equilibrium in the private information case.

Proposition 1. *In the case in which players do not know their capabilities, the unique symmetric BNE strategy is given by:* $s^*_{\varnothing} = \min\{1, \frac{1}{2e\int_{-\infty}^{\infty} f(c)^2 dc}\}$.

Remark. All proofs of propositions are presented in Appendix A. □

Now we turn to the disaster risk. This is the expected probability of disaster over the distribution of agents who play their BNE strategies. Since all agents are playing the same action, the expected risk of disaster is given by $1 - s^*_{\varnothing}$.

Corollary 1.1. *In any distribution of capability levels the expected level of disaster risk is given by* $D_{\varnothing} = \max\{0, 1 - \frac{1}{2e\int_{-\infty}^{\infty} f(c)^2 dc}\}$.

4.2 Private information

In this section, we consider the case in which each player knows her own capability level but not that of her rivals. This situation more closely resembles real-world qualitative races. Power- or profit-maximizing states and firms alike have a strong incentive to keep their technological capabilities a secret from rivals in order to win the technological race. Actors do not know when TAI will be developed and by whom, but they have some notion of how much progress they have made relative to the progress that their rivals are likely to have made. For instance, they may have developed beliefs derived from their research experience about whether TAI can be developed by scaling up computations, conditioning on other discoveries they have made, and how likely they and their rivals are to have access to required levels of computation on different time frames. Alongside other estimates, these factors lead to a set of subjective beliefs among rivals that are based on precise information about one's own achievements and guesses about the achievements of others. This situation is fairly well represented by the modeling scenario in which actors have private information about their capabilities. Thus, this scenario maps onto states of the world in which private firms or geopolitical rivals are developing TAI. In the private information case, each team i, can condition their safety level on their own capability, choosing a strategy: $s_{private}(x_i)$.[13] Given that players are maximizing expected utility, we use as our solution concept a Bayesian Nash equilibrium.

Proposition 2. *There always exists a symmetric Nash equilibrium in pure strategies. The strategy is given by* $s^*_{private}(x_i) = \min\{\frac{\int_{-\infty}^{x_i}(F_{n-1}(c))^e dc}{(F_{n-1}(x_i))^e}, 1\}$.

[13]Note that as our notation implies, we will focus on symmetric equilibria.

We see that as enmity level increases, the agents start putting less efforts into safety. Also, agents with higher capability exerts more effort into safety.

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* $D_{private} = 1 - 2 \cdot \int_0^{\mu} \min\{\frac{\int_{-\infty}^{x} F(c)^e dc}{F(x)^e}, 1\} \frac{F(x)}{\mu} dx$.

4.3 Public information

Now we solve for the case in which all agents are fully aware of each other's capabilities. This maps to states in which AI development remains open, as could be the case if academia is the driver of progress in the field, or geopolitical allies share information while developing TAI, or espionage techniques make secret-keeping impossible. Here, denote the leader's capability as x and the second highest player's as y. Denote $\Delta := x - y$ as the variable on which players condition their safety choices.

Consider the 2 player case. Each player's utility is given by

$$u_i(\Delta) = s_i(\Delta)\frac{e^{m(x_i - s_i)}}{\sum_{j=1}^{2} e^{m(x_j - s_j)}} + s_j(\Delta)(1-e)(1 - \frac{e^{m(x_i - s_i)}}{\sum_{j=1}^{2} e^{m(x_j - s_j)}})$$

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 $s^*(\Delta)$. We now show payoff and strategy equivalence with the Armstrong et al. [2016] model as $m \to \infty$.

Corollary 3.1. *[Strategy equivalence]*

$lim_{m\to\infty} s_x^*(\Delta) = min\{1, \frac{\Delta}{e}\}, \quad lim_{m\to\infty} s_y^*(\Delta) = (1-e)\frac{\Delta}{e}.$

Corollary 3.2. *[Payoff equivalence]*

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

5. Information hazards

In this section, we present comparative statics. Due to the complexity of the expression of the density function, we turn to numeric simulations to present our results.[14]

5.1 Information hazards

As in Armstrong et al. [2016], we are interested in the role of information in altering disaster risk. 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 1. 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. This is to be expected; the more expected value cutting corners in safety has in winning the race, the more likely players are to do so.[15]

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

[15] Though our statement for public information is relatively weaker, we note that in simulations of "reasonable" parameter values, such as those simulated in 1. risk sharply increases with decisiveness in the public information case as well, only mildly decreasing as m increases from 20 to ∞.

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 at* $m = \infty$ *than as* $m \to 0$.

The second set of results characterizes how the information environment interacts with the decisiveness parameter. As we shall see, the conclusions regarding information hazards in Armstrong et al. [2016] are dependent upon decisiveness. In the model, moving to a more open information state can increase the risk of disaster. This is what Bostrom [2011] refers to as an information hazard, or any "risk that arises from the dissemination or the potential dissemination of (true) information that may cause harm or enable some agent to cause harm." Consistent with Armstrong et al. [2016], we find that the no information state is always safer than the other two information states. We formally verify that these results are robust to m in the next two propositions.

Proposition 5. *The no information scenario is always safer than the private information scenario, unless risk under both scenarios is* 0.

Proposition 6. *The no information scenario is weakly safer than the public information scenario.*

The most interesting case, however, is the comparison between the public and private information states. Armstrong et al. [2016] find that when μ and e are large, the no information state is safest, followed by the private information state and then the public information state. However, we show that this result is dependent upon m. For large values of m, their results hold. However, as m begins to decline, we see in Figure 1 that the public information scenario becomes *safer* than the private information scenario. Finally, as m tends to 0, we

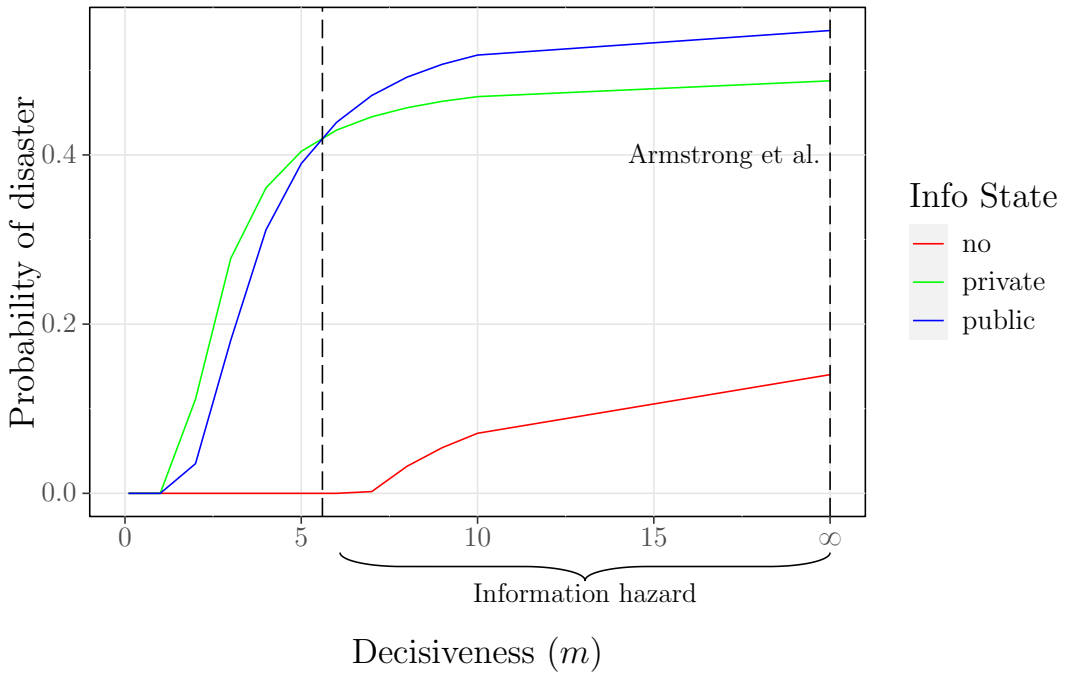

Figure 1: Disaster risk ($\mu = 1.44, e = 0.9$)

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 7. *The relative safety of the public and private information scenarios depends on m.*

Now we turn to the disaster risk. Consider Figure 1, where we see that the information hazard only obtains for $m > 6$. 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 $m \to \infty$ in Armstrong et al. [2016]. In the public information case, they show that when $n = 2$, the public information case is riskier than the private information

case as long as

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

In the public information case, risk is driven by cases in which the laggard is close behind the leader in capability; the probability of the leader and laggard being relatively close decreases linearly as μ, the upper bound on the capability-type distribution, increases.[16] In the private information case, risk is caused by low-capability winners. The probability that the winner has a relatively low capability decreases quadratically as μ increases. 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 would increase overall risk. However, players also see a lower expected return to reducing safety investments, which implies lower risk. In the public information state, when players are close in capability, they will continue to implement relatively risky strategies. In other cases, players know that they are far apart in capability and will increase safety relatively quickly as m falls. In contrast, in the private information state, players are always uncertain about the capability of their rivals. As a result, they never know that one player is effectively out of the running, and are thus less willing to increase safety investments either when their capabilities are relatively low or when they are relatively high - even when decisiveness is relatively low. In this state, the impact of m on the CSF is stronger. These effects are illustrated in Figure 2. In this figure, we simulate a race in which $\mu = 1, e = 0.9$. We fix $x_j = 0.5$ and consider what happens to expected

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

safety when we vary x_i. In the public information case, the race is most dangerous when $x_i = x_j = 0.5$. Both the leader and laggard cut corners in order to win, especially when decisiveness is high. However, in the private information case only low-capability players have a large incentive to cut corners. Though this varies with decisiveness, the rate at which safety decreases as decisiveness increases is considerably lower than in the public information case because 1) players never know they are close in capability, which is the condition that produces the dramatic increase in risk as decisiveness increases in the public information case, 2) low-capabilities types still have some uncertainty about whether they will indeed lose, and 3) the leader is fixed at $x_j = 0.5$, so the risky laggard is unlikely to win. Overall, then, these forces produce a result in which disaster risk declines relatively more quickly with m in the public information case than in the private information case, serving to reverse the information hazard in less decisive races.[17]

As stated in Propositions 5 and 6, the no information case is weakly safest. It turns out, then, that knowing one's own capability does not increase welfare. When actors do not, just as in cases with low decisiveness, players are quite uncertain about the returns to their own efforts. Cutting corners on safety is not worth it in these cases. In the no information case, with a uniform type distribution, the probability that players' capability levels are within δ of one another is $\left(\frac{\delta}{\mu}\right)^2$, and the chance of being a laggard is $\frac{1}{n}$. Players know their opponents are facing the same uncertainties. The combination of these two forces, therefore, means that the no information case is weakly safest even though risk also goes to zero when decisiveness is sufficiently low.

[17] In Appendix C, we show the size of the information hazard by plotting each difference in risk between information states.

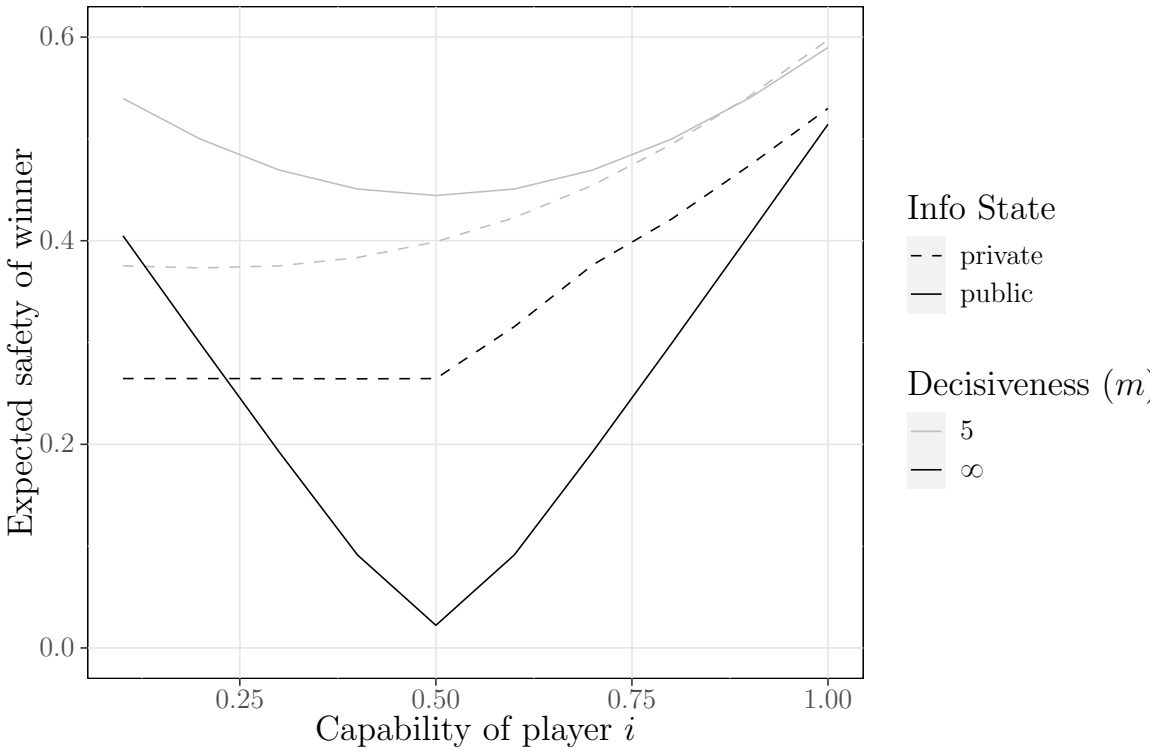

Figure 2: Expected safety of the race winner ($\mu = 1, e = 1, x_j = 0.5$)

6. Alternative safety provision function

This section lays out results when risk is determined by the choices of *both* the winner and the loser of the technology competition. As discussed in the introduction, we relax the extreme assumption of the Armstrong et al. [2016] model that the winner of the race is the sole source of risk, and allow the overall risk level to depend on the weighted sum of agents' safety efforts. This generalization mirrors real world races better, and raises a couple of 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 agents' incentive to win the race and the overall level of risk? As in the main section, we discuss the research question in the framework of a two agent game in each of the three information contexts (no information,

private information, and public information). For ease of reference, we reiterate equation (2), where γ is the relative impact of player i's choices on the overall level safety, given that player i is the competition winner.

$$\hat{s}_i = \gamma s_i + \frac{1-\gamma}{n-1}\sum_{j\neq i} s_j$$

6.1 No information case

In the no information case in which the agents are unaware of both the opponent's and her own endowed capability, the symmetric equilibrium dictates that all agents 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 8. *In the no information case, the equilibrium level of safety efforts in a symmetric BNE of pure strategies is given by:*

$$s^*_{\varnothing} = \min\{1, \frac{\gamma + (1-\gamma)(1-e)}{2e\mathbb{E}_{\boldsymbol{x}}\big(f(c)\big)}\}$$

The less safety is determined exclusively by the winner, the more the agents are willing to sacrifice safety to increase the chance of winning the race. In addition, we find that the amount by which the safety efforts decrease is linear in γ and is unaffected by the enmity level e. In other words, although the level of safety provision changes with the enmity level, the amount by which it changes with changing γ in the system is not dependent on the enmity level. Corollary 8.1 specifies the implied disaster risk.

Corollary 8.1. *In the no information case, the expected level of disaster risk is:*

$$D_{\varnothing} = \max\{0, 1 - \frac{\gamma + (1-\gamma)(1-e)}{2e\mathbb{E}\big(f(c)\big)}\}$$

6.2 Private information case

Under the private information scenario, the agents 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, as in the main section, the agents' strategy of how much to invest in safety will depend on their capabilities. Proposition 9 characterizes the equilibrium safety investments.

Proposition 9. *In the private information case, the equilibrium level of safety efforts in a symmetric BNE of pure strategies is given by:*

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

$$\textit{where } \Omega(x_i) = (1-e)(1-\gamma) + (\gamma - (1-e)(1-\gamma))F(x_i)$$

Unlike in Proposition 2, $s^*_{private}(x_i)$ is not an equilibrium for all $x_i \in supp(C)$.[18] The following two corollaries establish the uniqueness of $\underline{x}$ and the behavior of $x_i < \underline{x}$.

Corollary 9.1. *When $\gamma < 1$, if $\underline{x}$ exists, it is a convex set in $\mathbb{R}$.*

Corollary 9.2. *When $x_i < \underline{x}$, $s^*_{private}(x_i) = 1$.*

[18] Our proofs for Propositions 2 and 8 rely on the initial condition that $s^*(\underline{x}) = 0$. This cannot be a symmetric equilibrium for $e > 0$, as players have incentive to deviate to $s^*(\underline{x}) > 0$. Noting that $P[x_i = x] = 0$, we maintain that players hold this assumption in order to perform comparative statics with the model in Armstrong et al. [2016]. We leave it to future work to relax this assumption.

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. When $e = 0$, and the agents do not mind whether they or their opponent successfully builds the technology, Proposition 9 dictates that $s^*_{private} = 1$. When $e = 1$, and the agents are indifferent between disaster and victory for their adversaries, $s^*_{private} = \frac{\int F(x)^{\frac{1}{\gamma}} dx}{F(x)^{\frac{1}{\gamma}}}$. Remember from Proposition 2 that when $e = 1$, the equilibrium safety effort level was $\frac{\int F(x)dx}{F(x)}$. As γ falls, more equilibrium safety comes from the loser. This produces a *moral hazard* effect, in which highly capable actors choose lower safety levels because their returns to doing so are lower. 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[19]. This produces an additional *selection effect*. 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 actually makes agents put less effort into safety.

The calculation of the disaster risk under the private information case as well as the equilibrium outcome under the public information case is too involved to provide analytic solutions. Instead, we present simulated results and discuss them in the following subsection.

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

6.3 Public information case and simulation results

Figure 3 presents the disaster risk for $\gamma \in [0.5, 1]$ for the same parameters we used in simulations in Section 5 ($e = 0.9, \mu = 1.44$) 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.[20] Similar to the other cases, the same *selection effect* and *moral hazard* 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, similar to the moral hazard problem in standard public goods scenarios with weighted-sum provision technologies [Buchholz and Sandler, 2021]. Reduced investments in safety by others lowers the expected return on safety, even for actors that are likely to win the race.

7. Additional comparative statics

We now turn to comparative statics on the two remaining parameters of the model, enmity (e) and performance (μ).

[20] In the public information case for $\gamma \in [0, 0.5]$, risk is increasing in γ. However, we argue that this is unrepresentative of real-world scenarios, as it is unlikely that an actor who fails to implement TAI or other new technology contributes more risk than an actor who attempts implementation. Full results are presented in Appendix C.

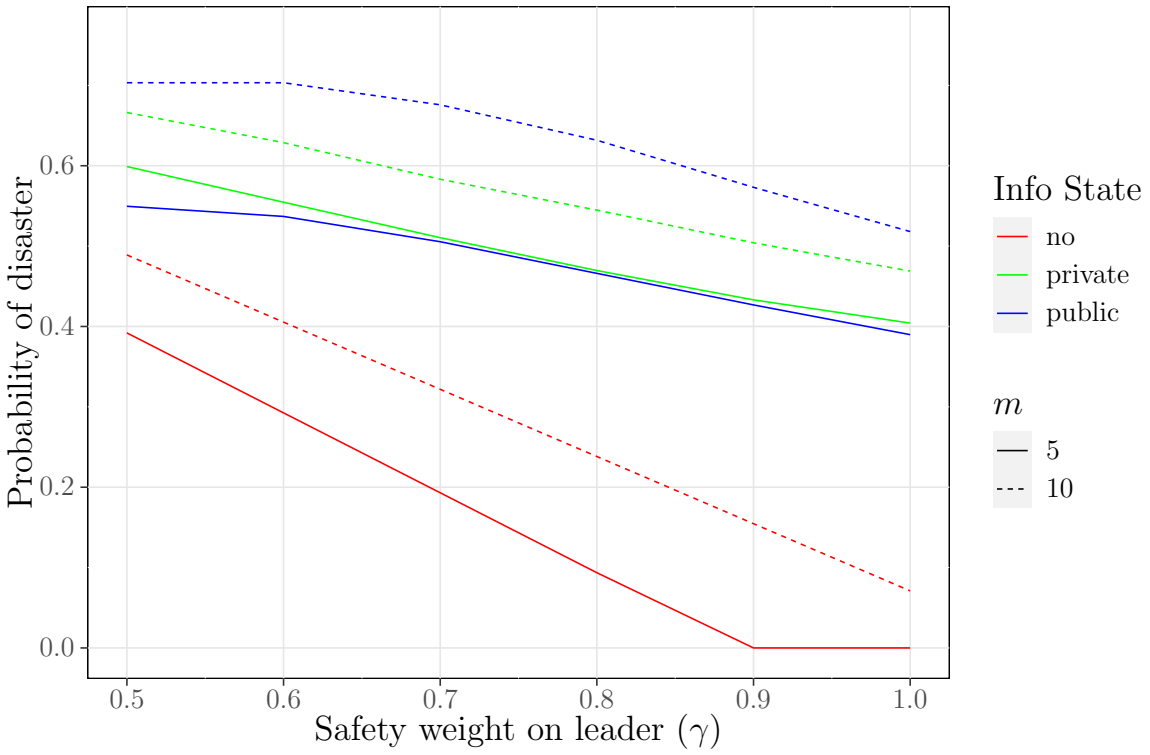

Figure 3: Varying safety contributions of the winner

7.1 Effects of enmity

The enmity parameter characterizes the opportunity cost of one's rival winning the race.[21] Increased enmity increases the opportunity cost of losing the race. When states are more intense rivals, they are more willing to cut corners to develop TAI. This is one of the reasons the United States government ignored some of Oppenheimer's concerns over the development of the atomic bomb: they feared Germany winning the race.[22] Presumably, the US would not have taken on the same level of risk had the UK been developing a nuclear bomb instead.

Results are presented for the public information case in Figure 4 for high and low values of enmity and of gamma. In all cases, increased enmity results in higher disaster risk. Higher

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

[22] Ord [2020].

values of m and lower values of γ increase the concavity of the curve. 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. Results for the no information and private information cases are similar and are presented in Appendix C. Qualitatively, then, the results in Armstrong et al. [2016] for enmity hold in this broader set of environments but these environments influence how harmful enmity is.

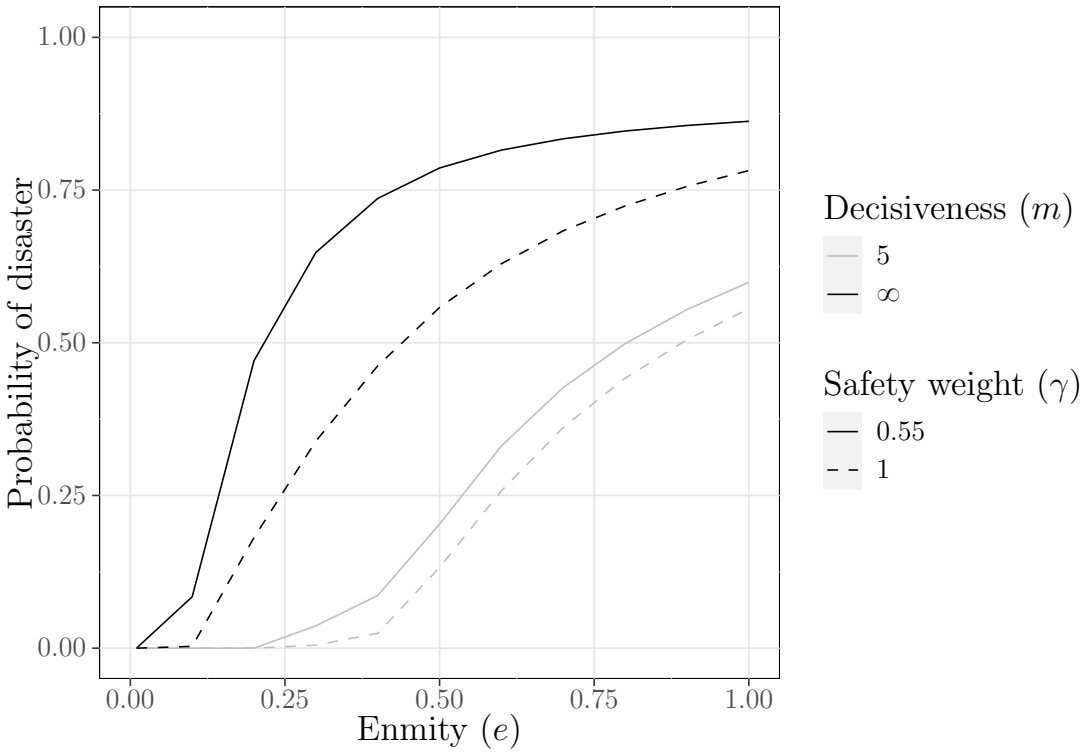

Figure 4: Effects of enmity under public information ($\mu = 0.72$)

7.2 Effects of type distribution

Next, we turn to analyzing the effects of performance, parametrized by μ.[23] In the model, μ represents the relative importance of performance to safety; if $\mu > 1$, performance is more important than cutting corners on safety to the outcome of the race. However, μ also represents the variance of players' performance levels. For smaller values of μ, players are more likely to be close together in performance, increasing the chance of a highly competitive race. In the Armstrong et al model, increasing μ always reduces the disaster risk, as $\frac{\partial Var(G(x_i))}{\partial \mu} = \frac{\partial}{\partial \mu}\frac{\mu^2}{12} = \frac{\mu}{6}$. We find that the magnitude of the effect is moderated by the decisiveness parameter. Results are presented in Appendix C. Though the risk is always nonincreasing in μ, the effect is larger for larger values of the decisiveness parameter m under all information states. Likewise, a larger value of γ increases risk for a fixed value of μ. In the no information and private information cases, a lower γ merely shifts the curve up, increasing risk at each value of μ. In the public information case, the curve is less steep - higher γ lessens the impact of μ. That is, in the public information case, the marginal effect of a higher variation in performance is lower when there is more safety weight on the loser than when the winner is the sole source of risk.

8. Discussion and conclusion

In the introduction, we suggested that the sources of risk in the Armstrong et al. [2016] model may not be representative of real-world technology races. The first factor that affects these

[23] We can generalize the effect of μ to consider other type distributions and their corresponding parameters. In Appendix B, we derive results on equilibrium safety levels as arbitrary type distributions are varied in the sense of first-order stochastic dominance.

results is the decisiveness parameter. In less decisive races, the information hazard vanishes, and sharing information about capabilities is welfare-improving. This is because, when the returns to risk-taking are relatively low, actors are reluctant to take risks if they realize they are far apart; only if actors come to realize they are close in capability will there be a race to the bottom. As a result, when forecasting the sources of risk for technology competitions, it remains important to research plausible paths through which TAI and other powerful technologies may be developed [Dafoe, 2017]. It is helpful to consider two cases, one in which the decisiveness parameter m is large or increasing without bound and another in which m is small. In the former case, the space of viable paths to TAI is relatively well-known. Currently, for example, we see large and relatively predictable gains from training deep learning models on ever larger amounts of data using increasing amounts of compute [Thompson et al., 2020]. If researchers realize that scaling up existing algorithms on larger training sets using ever larger numbers of GPUs will produce TAI, then differences in performance between leading teams may be highly decisive, particularly if transformative properties are expected to emerge suddenly once a certain scale is reached. In this case, according to the dynamics represented in the model, it is risk-reducing if AI teams or states are able to obscure their own performance to prevent dangerous races. In the latter case, the space of paths to TAI is relatively large. This maps to many real-world research scenarios, in which the space of possible ideas is increasing, driving down researcher productivity, as recent literature has shown to be the case [Agrawal et al., 2019a, Bloom, 2020].[24] Bostrom [2014] expresses concern about random, discontinuous progress in AI. Counterintuitively, from the point of

[24]This depends on the set of new general-purpose technologies growing at a slower rate than the decline in researcher productivity. If AI becomes a GPT before it becomes superintelligent [Agrawal et al., 2019a], then this trend may not hold.

view of a multiagent development race, we suggest that such a situation may actually be *safer* than a more deterministic production process since it will induce firms or states to take fewer risks.[25]

Whether the actions of actors who will not win the technology competition contribute to risk is a second important factor influencing overall risk from competition. We show that as the choices of losers become more impactful, the race becomes more dangerous. In these scenarios, actors face a more traditional Prisoner's Dilemma scenario, in which each is tempted to shirk on safety provision because they do not bear the full benefits of their safety investments yet absorb all of the costs. To the extent that race losers contribute to risk, a multilateral regulatory regime may be required that takes into account the greater willingness of race laggards to cut corners on safety.

We hope that the findings of our model can be extended in a number of directions and therefore suggest directions for future research. First, we assume that the information partitions are given exogenously. In the real world, however, actors may choose to share information or close off development to increase their own chances of winning the race. Technology firms often face a dilemma: research publications and capability demonstrations garner prestige and attention that can be necessary to raise funds for further operations; yet, they also aid the efforts of competitors. An important extension would be to endogenize these decisions to understand how they affect disaster risk. This would assist actors maing decisions about how open to be, for instance about the publishing of source code for advanced AI models. OpenAI faced this question with its GPT-3 language model, deciding against

[25]Note that this model still assumes that actors always know what "AI safety" looks like and can therefore optimize for it. When the race is more noisy, perhaps one can expect the productivity of certain types of safety research to be more noisy as well.

publication, but Meta recently released the source code of a model designed to replicate GPT-3.[26] In addition, we can analyze scenarios in which actors can partially reveal information or uncover information about rivals, since this likely mirrors real world dynamics between firms and states. 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 actors choose higher safety levels when the tradeoff is more concave. It is likely to be particularly important to develop models with an explicit price of safety which is a function of the level of performance. Lower returns to safety relative to performance could also be a result of uncertainty. For a risk averse agent, 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 the innovation literature suggests. Therefore, extending the model to a dynamic game in which decisiveness varies over the course of the race might yield insights. 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 actors to increase general welfare by conditioning their strategies on each other's behavior [Stafford and Trager, 2022].

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, when combined with others, are not only theoretically useful but can be used to improve policy decisions so that advanced technologies are developed for the benefit of all.

[26] See Zhang and Diab.

Appendix A: Proofs

On the definition of the random variable C_i. First, we note that a logistic contest success function corresponds to Gumbel-distributed noise [Ryvkin and Drugov, 2020, McFadden, 1974]. Therefore, we can represent each i's type with the random variable $C_i = X_i + V_i - V_j$, where $X_i \sim Uniform(0, \mu)$ and $V_i, V_j \sim Gumbel(1, \frac{1}{m})$, where m is the decisiveness parameter.

$$F_{n-1}(c) = Pr(\text{all } c_i < c)$$

Now, conditional on v_j, c_i are independent for all i. We now solve for $F(c)$:

$$F(c) = Pr(c_i < c|v_j) = Pr(x_i + v_i - v_j < c)$$

Finally, we have

$$F_{n-1}(c) = F^{n-1}(c)$$

Next, to simplify calculations, we show that $V_i - V_j \sim Logistic(0, \frac{1}{m})$. To prove that these distributions are identical, it suffices to show that the moment-generating functions are identical [Hogg et al., 2012].

The MGF of two independent $Gumbel(1, \frac{1}{m})$-distributed random variables is

$$M_{V_i - V_j}(t) = M_{V_i}(t) M_{V_j}(-t) = \Gamma(1 - \frac{1}{m}t)\mathrm{e}^{t}\Gamma(1 + \frac{1}{m}t)\mathrm{e}^{-t} = \Gamma(1 - \frac{1}{m}t)\Gamma(1 + \frac{1}{m}t)$$

The MGF of a $Logistic(0, \frac{1}{m})$-distributed random variable is

$$M(t) = \frac{\Gamma(1 - \frac{1}{m}t)\Gamma(1 + \frac{1}{m}t)}{\Gamma(2)} \mathrm{e}^{0*t} = \Gamma(1 - \frac{1}{m}t)\Gamma(1 + \frac{1}{m}t) = M_{V_i - V_j}(t) \quad \forall t \in (-m, m)$$

Therefore, the random variable of interest becomes $C_i = X_i + Y_i$, where $Y_i \equiv V_i - V_j \sim Logistic(0, \frac{1}{m})$. Now the problem simplifies to calculating $F(c)$. The CDF of the sum of independent random variables $W = U + Y$ is given by the convolution

$$F_w(w) = (F_U * F_Y)(w) = \int F_U(w - y) f_Y(y) dy$$

Substituting in our random variables, we have

$$F_C(c) = (F_X * F_Y)(c) = \int F_X(c - x) f_C(x) dx = \frac{1}{\mu} \int_0^{\mu} \frac{1}{1 + \mathrm{e}^{-m(c-x-1)}} dx$$

To complete the proof, recall that $F_{n-1}(c) = F^{n-1}(c)$.

Finally, to calculate $f(c)$, we just take the derivative of $F(c)$ with respect to c, giving

$$f(c) = \frac{1}{\mu} \int_0^{\mu} \frac{m \cdot \mathrm{e}^{-m(c-x-1)}}{(1 + \mathrm{e}^{-m(c-x-1)})^2} dx$$

□

Proof of Proposition 1. The utility of player i is given by

$$\begin{aligned} u_i(s_i|s_{-i}) &= Pr(i \text{ wins})s_i + (1-e)\sum_{k\neq i} Pr(k \text{ wins})s_k \\ &= (1 - \sum_{k\neq i} Pr(k \text{ wins}))s_i + (1-e)\sum_{k\neq i} Pr(k \text{ wins})s_k \\ &= s_i + \sum_{k\neq i} Pr(k \text{ wins})((1-e)s_k - s_i) \end{aligned}$$

Using $\mathbb{E}_k$ as the expectation over the realizations of c_k and by the law of iterated expectations, we get the following expression for expected utility:

$$\begin{aligned} u_i(s_i) &= s_i + \mathbb{E}\Big(\sum_{k\neq i} Pr(k \text{ wins})((1-e)s_k - s_i)\Big) \\ &= s_i + \sum_{k\neq i} \mathbb{E}_k(Pr(k \text{ wins}|c_k))((1-e)s_k - s_i) \end{aligned}$$

From the definition of C_i,

$$\begin{aligned} Pr(k \text{ wins}|c_k) &= Pr(c_j \leqslant c_k - s_k + s_j \quad \forall j|c_k) \\ &= Pr(x_j + v_j \leqslant c_k - s_k + s_j \quad \forall j|v_i, c_k) \\ &= Pr(x_j + v_j - v_i \leqslant c_k - s_k + s_j \quad \forall j|c_k) \\ &= Pr(x_j + v_j - v_i \leqslant c_k \quad \forall j|c_k) = F_j(c_k) \end{aligned}$$

Where we exploit conditional independence of v_j's on v_i, and the symmetry of strategies.

Plugging into our utility function gives

$$u_i(s_i) = s_i + \sum_{k \neq i} \mathbb{E}_k[F_{n-1}(c_k)]((1-e)s_k - s_i)$$

Taking the first order conditions gives

$$\frac{\partial u_i(s_i)}{\partial s_i} = 1 - \sum_{k \neq i} \mathbb{E}_k[F_{n-1}(c_k)] + \sum_{k \neq i} \mathbb{E}_k[F_{n-2}(c_k)f(c_k)]((1-e)s_k - s_i) = 0 \tag{4}$$

We again use symmetry to obtain

$$1 - (n-1)\mathbb{E}[F_{n-1}(c)] - es(n-1)\mathbb{E}[f(c)F_{n-2}(c)] = 0$$

Recalling that we are considering the case where $n = 2$, we have

$$s_i^* = \frac{1 - \mathbb{E}[F(c)]}{2e\mathbb{E}[f(c)]}$$

Note that since $s_i = s_{-i}^* \equiv s_\varnothing$, we have $\mathbb{E}[F(c)] = \frac{1}{2}$. Recalling that $s_\varnothing^* \in [0, 1]$, we have

$$s_{\varnothing^*} = min\{\frac{1}{2e\mathbb{E}[f(c)]}, 1\}$$

Now we show that this expression is a local maximum for all parameter values. Taking the derivative of (4), we have

$$\frac{\partial^2 u_i(s_i)}{\partial s_i^2} = -2\sum_{k\neq i}\mathbb{E}_k[f_i(c_k)F_{n-2}(c_k)] + \sum_{k\neq i}\mathbb{E}_k[\frac{\partial}{\partial s}f_i(c_k)F_{n-2}(c_k)](-es)$$

Exploiting the symmetry of strategies, we have

$$\frac{\partial^2 u_i(s_i)}{\partial s_i^2} = -2(n-1)\mathbb{E}[f(c)F_{n-2}(c)] - es(n-1)\mathbb{E}[f'(c)F_{n-2}(c)] \tag{5}$$

Evaluating the above at $s^*_\varnothing$ obtained in proposition 1 and considering the case when $n = 2$, we find that $s^*_\varnothing$ is a local symmetric BNE when the following holds:

$$\frac{\partial^2 u_i(s_i)}{\partial s_i^2} = -2\int_{-\infty}^{\infty} f(c)^2 dc - \frac{\int_{-\infty}^{\infty} f'(c)f(c)dc}{2\int_{-\infty}^{\infty} f(c)^2 dc} \leqslant 0$$

Exploiting the symmetry of the PDF in the definition of C_i and using integration by parts, we have

$$\mathbb{E}[f'(c)] = \frac{f^2(c)}{2}\Big|_{-\infty}^{\infty} = 0$$

Therefore, we can simplify the second order condition as

$$\frac{\partial^2 u_i(s_i)}{\partial s_i^2} = -2\underbrace{\int_{-\infty}^{\infty} f(c)^2 dc}_{\geqslant 0} \leqslant 0$$

This expression always holds.

□

Proof of Corollary 1.1. Since all players are implementing the same safety level $s^*_{\varnothing}$ given by Proposition 1, the total expected probability of an AI disaster is $1 - s^*_{\varnothing} = max\{0, 1 - \frac{1}{2e\int_{-\infty}^{\infty} f(c)^2 dc}\}$. □

Proof of Proposition 2. We prove the existence by construction. Define $k(x_i) = x_i - s(x_i)$ as our choice variable. We assume and confirm that $k(\cdot)$ is an increasing function, which means that the agent with higher innate capability ends up performing better.

Then the expected utility of player i is given by

$$u_i(k(x_i)) = x_i - k(x_i) + \int_{k^{-1}(k(x_i))}^{\infty} ((1-e)(c-k(c)) - (x_i - k(x_i)))f_{n-1}(c)dc$$

Taking the first-order conditions gives

$$\frac{\partial u_i(k(x_i))}{\partial k(x_i)} = -1 + \frac{\partial}{\partial k(x_i)} \int_{k^{-1}(k(x_i))}^{\infty} ((1-e)(c-k(c)) - (x_i - k(x_i)))f_{n-1}(c)dc$$

Now we evaluate the integral term.

Note that $\int_{k^{-1}(k(x_i))}^{\infty} ((1-e)(c-k(c)) - (x_i - k(x_i)))f_{n-1}(c)dc = I(\infty) - I(k^{-1}(k(x_i)))$, where $I(c)$ is the anti-derivative of the integrand with respect to c.

Then we have

$$\begin{aligned}
&\frac{\partial}{\partial k(x_i)} \int_{k^{-1}(k(x_i))}^{\infty} ((1-e)(c-k(c)) - (x_i - k(x_i)))f_{n-1}(c)dc \\
&\qquad = \frac{\partial}{\partial k(x_i)} I(\infty) - \frac{\partial}{\partial k(x_i)} I(k^{-1}(k(x_i))) \\
&\qquad = \frac{\partial}{\partial k(x_i)} k(x_i) F_{n-1}(\infty) - \frac{\partial}{\partial k(x_i)} I(k^{-1}(k(x_i))) \\
&\qquad = F_{n-1}(\infty) - \frac{\partial}{\partial k(x_i)} I(k^{-1}(k(x_i)))
\end{aligned}$$

Note that $F_{n-1}(\infty) = 1$. Using the chain rule, we have

$$\begin{aligned}
&1 - \frac{\partial}{\partial k(x_i)} I(k^{-1}(k(x_i))) \\
&= 1 - F_{n-1}(k^{-1}(k(x_i))) - \Big[(1-e)(c-k(c) - (x_i - k(x_i)))f_{n-1}(c)\Big]_{c=k^{-1}(k(x_i))} \cdot \frac{\partial}{\partial k(x_i)} k^{-1}(k(x_i))
\end{aligned}$$

By the implicit function theorem, we know $\frac{\partial}{\partial k(x_i)} k^{-1}(k(x_i)) = \frac{1}{k'(k^{-1}(k(x_i)))}$, where $k' := k'(c) = 1 - s'(c)$.

Plugging back into the first-order conditions gives

$$\begin{aligned}
&\frac{\partial u_i(k(x_i))}{\partial k(x_i)} = 0 \\
&\qquad = -F_{n-1}(k^{-1}(k(x_i))) \\
&\qquad\quad - [(1-e)(k^{-1}(k(x_i)) - k(x_i)) - (x_i - k(x_i))] f_{n-1}(k^{-1}(k(x_i))) \frac{1}{k'(k^{-1}(k(x_i)))}
\end{aligned}$$

Applying symmetry of strategies gives $k(x_i) = k_j(x_i) \equiv k(x_i)$. Therefore, $k^{-1}(k_i(x_i)) = k^{-1}(k(x_i)) = x_i$.

This allows us to simplify to

$$\frac{\partial u_i(k(x_i))}{\partial k(x_i)} = 0 = -F_{n-1}(x_i) + e(x_i - k(x_i))f_{n-1}(x_i)\frac{1}{k'(x_i)} \tag{6}$$

We know that a general linear ordinary differential equation has the form $y'(x) + p(x)y(x) = q(x)$, with solution $y(x) = \frac{1}{r(x)}\int r(x)q(x)dx + \frac{Constant}{r(x)}$, where $r(x) = \mathrm{e}^{\int p(x)dx}$.

We can therefore express our FOC ODE as:

$$k(x_i) = \frac{1}{r(x_i)}\int_{\underline{x}}^{x} r(x_i)q(x_i)dx_i + \frac{\text{const}}{r(x_i)}$$

$$\text{where} \quad r(x_i) = e^{\int e\frac{f_{n-1}(x_i)}{F_{n-1}(x_i)}dx_i} = (m\mu F_{n-1}(x_i))^e$$

$$\Rightarrow \quad k(x_i) = \frac{1}{(F_{n-1}(x_i))^e}\int_{\underline{x}}^{x_i} ec(F_{n-1}(c))^{e-1}f_c(x_i)dc + \frac{\text{const}}{(m\mu F_{n-1}(x))^e} + lim_{c\to\underline{x}}(k(c)-c)\frac{(m\mu F_{n-1}(c))^e}{(F_{n-1}(x_i))^e}$$

$$= \frac{1}{(F_{n-1}(x_i))^e}\big((F_{n-1}(x_i))^e \cdot x_i - \int_{\underline{x}}^{x_i}(F_{n-1}(c))^e dc\big) + \frac{\text{const}}{(m\mu F_{n-1}(x))^e}$$

$$+ lim_{c\to\underline{x}}(k(c)-c)\frac{(m\mu F_{n-1}(c))^e}{(F_{n-1}(x_i))^e}$$

where we mapped $(x, y(x), p(x), q(x))$ to $(x_i, k(x_i), e\frac{f_{n-1}(x_i)}{F_{n-1}(x_i)}, e\frac{f_{n-1}(x)}{F_{n-1}(x)}x)$.

Bringing the notation back to $s(x_i)$ terms again:

$$s(x_i) = \frac{\int_{-1}^{x_i}(F_{n-1}(c))^e dc}{(F_{n-1}(x_i))^e} + \frac{\text{const}}{(m\mu F_{n-1}(x))^e} + lim_{c\to\underline{x}}s(c)\frac{(m\mu F_{n-1}(c))^e}{(F_{n-1}(x_i))^e}$$

Where we note that the initial condition $s(\underline{x}) = 0$ gives that the constant term goes to 0. Note that the ODE must hold for all $x_i \in [\underline{x}, \mu]$. Otherwise, the FOC fail to produce a local maximum. We conjecture that $\underline{x} = -\infty$ and verify this using the SOC.

Finally, recall that $s_i(x_i) \in [0, 1]$. Therefore, the equilibrium symmetric BNE in the two player game is given by

$$s^*_{private}(x_i) = min\{\frac{\int_{-\infty}^{x_i}(F_{n-1}(c))^e dc}{(F_{n-1}(x_i))^e}, 1\}$$

To check that such local optimum is indeed a maximum, we check the second order condition:

$$\frac{\partial^2 u_i(k_i)}{\partial k_i^2} = -\frac{\partial P(k_i)}{\partial k_i} + ex_i\frac{\partial^2 P(k_i)}{\partial k_i^2} - e\frac{\partial P(k_i)}{\partial k_i} - ek_i\frac{\partial^2 P(k_i)}{\partial k_i^2}$$

Where we define $P(k(x_i)) := F(c_i)$. The second order condition therefore boils down to

$$s^*(x)\frac{f'(x)}{f(x)} \leqslant 1 + \frac{1}{e}$$

Which always holds for the given $s^*(x)$ when $x \sim$ Uniform distribution. □

Proof of Corollary 2.1. Letting $G(x)$ denote the CDF of x_j, we calculate the expected value of the winner's chosen safety efforts by multiplying for each agent: $g(x)$ – the probability that an agent is given performance x, $\mathbb{P}$(he wins), and $s(x)$ – the safety efforts chosen by that agent.

$$D_{private} = 1 - 2 \cdot \int_0^{\mu} s(x) \cdot \mathbb{P}(\text{x is the winner}) \cdot g(x)dx$$

Since we assumed in Proposition 2 that k_i is an increasing function of x, the probability that an agent with capability x wins the race is simply $F(x)$.

$$D_{private} = 1 - 2 \cdot \int_0^{\mu} s(x)F(x)\frac{1}{\mu}dx$$

□

Proof of Proposition 3. Denote the difference between the highest capability level, x, and the second-highest, y, as $\Delta \equiv x - y$.

The utility functions are given by

$$u_x(\Delta) = s_x \frac{1}{1 + \mathrm{e}^{m(-\Delta + s_x - s_y)}} + (1 - e)s_y(1 - \frac{1}{1 + \mathrm{e}^{m(-\Delta + s_x - s_y)}})$$
$$u_y(\Delta) = s_y \frac{1}{1 + \mathrm{e}^{m(\Delta + s_y - s_x)}} + (1 - e)s_x(1 - \frac{1}{1 + \mathrm{e}^{m(\Delta + s_y - s_x)}})$$

Taking first order conditions, we obtain

$$\frac{\partial u_x(\Delta)}{\partial s_x} = \frac{1}{1 + \mathrm{e}^{m(-\Delta + s_x - s_y)}} - s_x \frac{m\mathrm{e}^{m(-\Delta + s_x - s_y)}}{(1 + \mathrm{e}^{m(-\Delta + s_x - s_y)})^2} + (1 - e)s_y \frac{m\mathrm{e}^{m(-\Delta + s_x - s_y)}}{(1 + \mathrm{e}^{m(-\Delta + s_x - s_y)})^2}$$
$$\frac{\partial u_y(\Delta)}{\partial s_y} = \frac{1}{1 + \mathrm{e}^{m(\Delta + s_y - s_x)}} - s_y \frac{m\mathrm{e}^{m(\Delta + s_y - s_x)}}{(1 + \mathrm{e}^{m(\Delta + s_y - s_x)})^2} + (1 - e)s_x \frac{m\mathrm{e}^{m(\Delta + s_y - s_x)}}{(1 + \mathrm{e}^{m(\Delta + s_y - s_x)})^2}$$

Setting these equations equal to 0 and simplifying, we obtain

$$1 + (1 - ms_x + m(1-e)s_y)\mathrm{e}^{m(-\Delta+s_x-s_y)} = 0$$

$$1 + (1 - ms_y + m(1-e)s_x)\mathrm{e}^{m(\Delta+s_y-s_x)} = 0$$

We solve the above implicitly for $s_y^*(\Delta), s_x^*(\Delta)$, respectively, to obtain

$$s_y^*(\Delta) = \frac{(1-e)W(-\frac{\mathrm{e}^{m\Delta-ms_x-ms_x/(e-1)+1/(e-1)}}{e-1}) - ms_x + 1}{m(e-1)}$$
$$s_x^*(\Delta) = \frac{(1-e)W(-\frac{\mathrm{e}^{-m\Delta-ms_y-ms_y/(e-1)+1/(e-1)}}{e-1}) - ms_y + 1}{m(e-1)}$$

where $W(.)$ represents branch 0 of the Lambert W function. □

Proof of Corollary 3.1. We show that $s_y^*(\Delta), s_x^*(\Delta)$ converge to their values in Armstrong et al. [2016] as $m \to \infty$. We first take the limit of $s_y^*(\Delta; m)$ as $m \to \infty$. Using l'Hopital's rule, we obtain

$$\begin{aligned}
lim_{m\to\infty} s_y(\Delta) &= lim_{m\to\infty} \frac{(1-e)W(-\frac{\mathrm{e}^{m\Delta-ms_x-ms_x/(e-1)+1/(e-1)}}{e-1}) - ms_x + 1}{m(e-1)} \\
&= lim_{m\to\infty} \frac{(1-e)(\Delta - s_x - \frac{s_x}{e-1})\frac{W(-\frac{\mathrm{e}^{m\Delta-ms_x-ms_x/(e-1)+1/(e-1)}}{e-1})}{W(-\frac{\mathrm{e}^{m\Delta-ms_x-ms_x/(e-1)+1/(e-1)}}{e-1})+1} - s_x}{e-1} \\
&= \frac{-s_x + (1-e)(\Delta - s_x - \frac{s_x}{e-1})}{e-1} \\
&= s_x - \Delta
\end{aligned}$$

We plug this value into the first-order condition for s_y to obtain

$$lim_{m\to\infty}\frac{1}{m} + (\frac{1}{m} - s_y + (1-e)s_x)\mathrm{e}^{m(\Delta+s_y-s_x)} = lim_{m\to\infty}0$$

$$s_x - \Delta + (1-e)s_x = 0$$

$$s_x(\Delta) = \frac{\Delta}{e}$$

This is the value of $s_x^*(\Delta)$ given in Armstrong et al. [2016]. To see that player y's strategy converges, note that player x chooses s_x such that y is indifferent between tying x and abstaining: $x - s_x = y - s_y \Rightarrow s_y = s_x - \Delta$. □

Proof of Corollary 3.2. From Tullock [2005], we know that

$$lim_{m\to\infty}\frac{\mathrm{e}^{m(x_i-s_i)}}{\sum_{j=1}^{2}\mathrm{e}^{m(x_j-s_j)}} = \begin{cases} 1 & \text{if } x_i > x_j \\ 0 & \text{if } x_i < x_j \end{cases}$$

Plugging this limit back into each player's utility functions completes the proof. □

Proof of Proposition 4. To simplify notation, denote F_a as the derivative of the distribution function F with respect to parameter a. First, consider the no information case. Differenti-

ating the unbounded version of the safety expression in 1 with respect to m yields

$$\begin{aligned}\frac{\partial s_{\varnothing}}{\partial m} &= \frac{-2e\int 2f(c)f_m(c)dc}{4e^2\mathbb{E}[f(c)]^2} \\ &= \frac{-\int 2f(c)f_m(c)dc}{2e\mathbb{E}[f(c)]^2}\end{aligned}$$

We note that all of the terms are positive expect perhaps $f_m(c)$. Differentiating $f(c)$ pointwise, we have

$$\frac{\partial f}{\partial m} = \frac{1}{\mu}\left(2\mu + (1-\frac{1}{e^{2cm}})(\frac{ce^{-cm}}{1+e^{-2cm}}) + (1-\frac{1}{e^{m(\mu-c)}})(\frac{(\mu-c)e^{m(\mu-c)}}{1+e^{2m(\mu-c)}})\right) > 0$$

Thus, plugging this expression into our derivative and adding bounds

$$\frac{\partial s_{\varnothing}}{\partial m} \leqslant 0$$

Next, consider the private information case. We take the derivative of the unbounded version of the safety expression in 2.

$$\begin{aligned}\frac{\partial s_{private}(x;m)}{\partial m} &= \frac{(F(x)^e\frac{\partial}{\partial m}\int F(c)^e dc) - eF(x)^{e-1}F_m(x)\int F(c)^e dc}{F(x)^{2e}} \\ &= \frac{\frac{\partial}{\partial m}\int F(c)^e dc - eF(x)^{-1}F_m(x)\int F(c)^e dc}{F(x)^e} < 0\end{aligned}$$

Adding in bounds, we have

$$\frac{\partial s_{private}(x;m)}{\partial m} \leqslant 0$$

Finally, we consider the public information case. We consider the limit of $s_x(\Delta), s_y(\Delta)$ as m goes to 0.

$$lim_{m\to 0} s_y(\Delta) = lim_{m\to 0} \frac{(1-e)W(-\frac{e^{-m\Delta - ms_y - ms_y/(e-1)+1/(e-1)}}{e-1}) - ms_y + 1}{m(e-1)} = 1$$
$$lim_{m\to 0} s_y(\Delta) = lim_{m\to 0} \frac{(1-e)W(-\frac{e^{m\Delta - ms_x - ms_x/(e-1)+1/(e-1)}}{e-1}) - ms_x + 1}{m(e-1)} = 1$$

Likewise, we use the expressions for safety when $m = \infty$ in 3.1 to note that

$$\mathbb{P}(D_{public} = 1) = 1 - \mathbb{P}(s_x\Delta = 0 \cap s_y\Delta = 1) = 1 - \mathbb{P}(\Delta = 0) = 0$$

Thus, we conclude that

$$\mathbb{E}[s_{public}(\Delta; m = \infty)] > \mathbb{E}[s_{public}(\Delta; m = 0)]$$

□

Proof of Proposition 5. We first establish that $s_{private}(x_i)$ is a weakly increasing function of x_i. Taking the derivative of the safety expression in Proposition 2 with respect to x_i yields:

$$\frac{\partial s_{private}(x_i)}{\partial x_i} = \frac{F(x_i)^{2e} - e\int F(x_i)dx_i F(x_i)^{e-1} f(x_i)}{F(x_i)^{2e}}$$
$$= 1 - s_{private}(x_i)\frac{ef(x_i)}{F(x_i)} \geqslant 0$$

Since $s_{private}(x_i)$ is non-decreasing in x_i, if $s_\varnothing > s_{private}(\mu)$, it automatically follows that $s_\varnothing > s_{private}(x)$ for any x.

Since $s_{\varnothing}$ is constant for all x_i, it remains to show that

$$s_{\varnothing} - s_{private}(\mu) = \frac{1}{2e\mathbb{E}[f(c)]} - \frac{\int_{-\infty}^{\mu} F(x_i)^e dx_i}{F(\mu)^e} \geqslant 0$$

Calculations reveal that this expression always holds for $m > 0$.

Note that while disaster risk under the no information scenario is constant at $1 - s_{\varnothing}$, disaster risk under the private information is bounded below by $1 - s_{private}(\mu)$, which, by definition, occurs when all parties involved put in maximal safety efforts. Therefore, the above exposition that $s_{\varnothing} > s_{private}(\mu)$ under all parameters completes the proof that no information scenario is always safer than the private information, regardless of the players' position along x. □

Proof of Proposition 6. Because of the complexity of the implicit function defining $s^*_{public}(\Delta)$, here we provide a graphical proof sketch to demonstrate that

$$\mathbb{E}[s^*_{\varnothing}] \geqslant \mathbb{E}[s^*_{public}(\Delta)]$$

holds for all values of e, μ, m. To do this, consider first extreme values of enmity ($e = \{0, 0.5, 1\}$) and a typical value of $\mu = 1.44$. Figure 5 plots the results. Note that $\mathbb{E}[s^*_{\varnothing} - s^*_{public}(\Delta)] \geqslant 0 \ \forall m$.

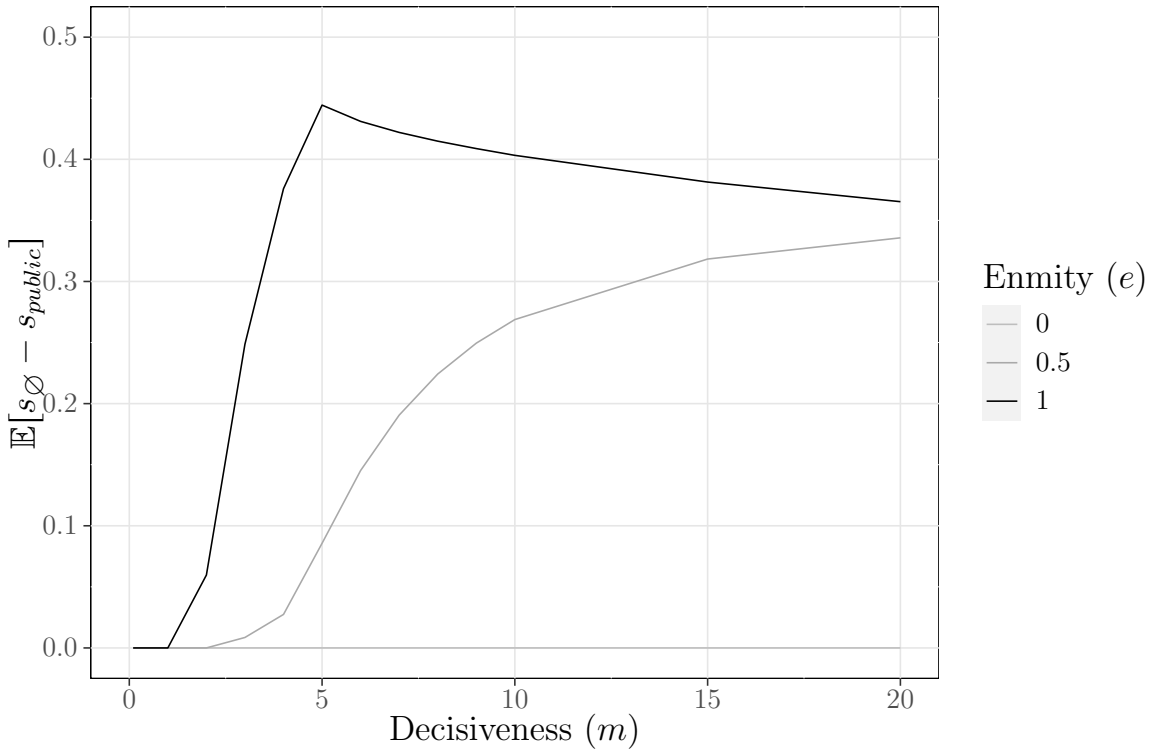

Figure 5: Extreme values of enmity ($\mu = 1.44$)

Next, consider extreme values of the performance parameter ($\mu = \{0.01, 12.5, 25\}$) and a typical value of $e = 0.9$. Figure 6 plots the results. Here, too, $\mathbb{E}[s^*_\varnothing - s^*_{public}(\Delta)] \geqslant 0$ holds for all m.

Given that expected safety under no information is always at least as large as under public information for these extreme values of the parameter space, we can be reasonably certain that this result holds for the entire parameter space.

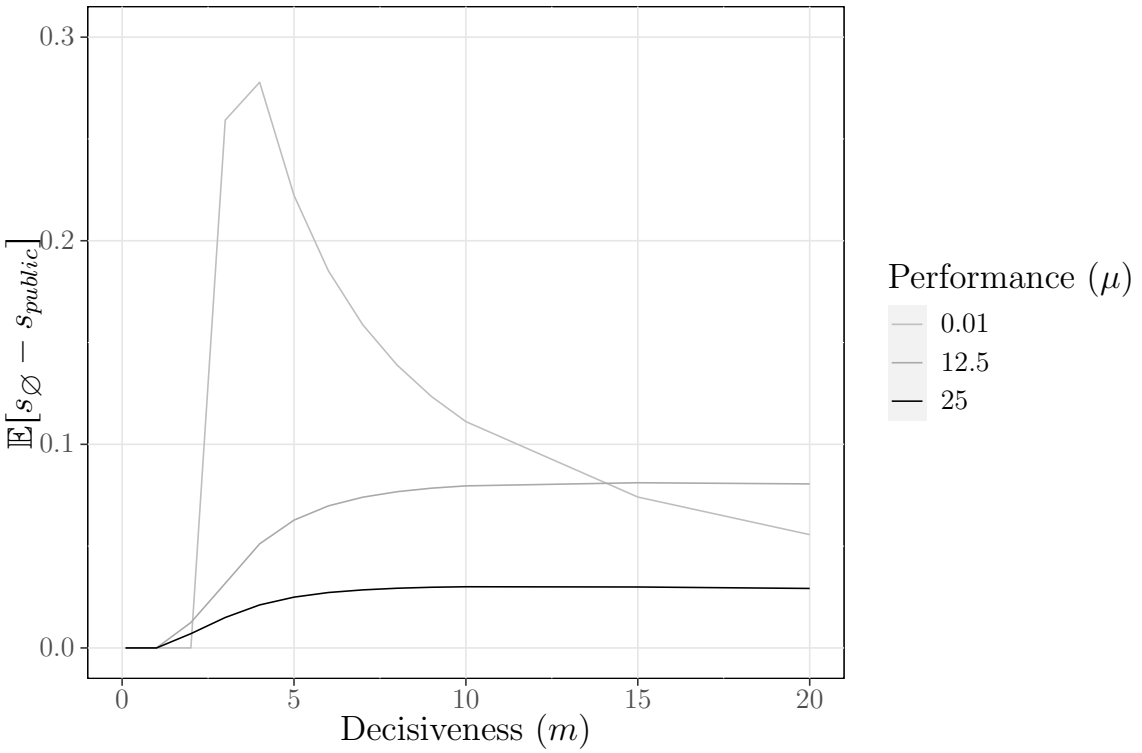

Figure 6: Extreme values of performance ($e = 0.9$)

□

Proof of Proposition 7. A numeric example from Figure 1 suffices for this proof. Let $\mu = 1.44, e = 0.2, n = 2$. Notice from the following two cases that the relative risk between private and the public information scenarios depend on the parameters.

$$\mathbb{E}[s_{private}(x; m = 4)] = 0.6389 < \mathbb{E}[s_{public}(\Delta; m = 4)] = 0.6885$$

$$\mathbb{E}[s_{private}(x; m = 7)] = 0.5549 > \mathbb{E}[s_{public}(\Delta; m = 7)] = 0.5297$$

□

Proof of Proposition 8 and Corollary 8.1. Denoting $p_i(\boldsymbol{s}, \boldsymbol{x})$ as probability that i wins given strategy profile $\boldsymbol{s}$ and capability vector $\boldsymbol{x}$, we get:

$$\begin{aligned} u_i(s_i|s_j, \boldsymbol{x}) &= (1 - p_j(\boldsymbol{s}, \boldsymbol{x})) \cdot \hat{s_i} + (1 - e) \cdot p_j(\boldsymbol{s}, \boldsymbol{x}) \cdot \hat{s_j} \\ &= \hat{s_i} + p_j(\boldsymbol{s}, \boldsymbol{x}) \cdot ((1 - e)\hat{s_j} - \hat{s_i}) \\ &= \gamma s_i + (1 - \gamma) s_j \\ &\quad + p_j(\boldsymbol{s}, \boldsymbol{x}) \cdot \big(((1 - e)(1 - \gamma) - \gamma) s_i - ((1 - e)\gamma - (1 - \gamma)) s_j\big) \end{aligned}$$

Taking the FOC with respect to s_i yields:

$$\begin{aligned} \frac{\partial u_i(s_i|s_j, \boldsymbol{x})}{\partial s_i} &= \gamma + ((1 - e)(1 - \gamma) - \gamma) p_j(\boldsymbol{s}, \boldsymbol{x}) \\ &\quad + \frac{\partial p_j(\boldsymbol{s}, \boldsymbol{x})}{\partial s_i} \big(\underbrace{((1 - e)(1 - \gamma) - \gamma) s_i - ((1 - e)\gamma - (1 - \gamma)) s_j}_{-es^*} \big) = 0 \end{aligned}$$

Finally, we utilize symmetry of strategies $s_i = s_j = s^*$, and of winning probabilities $\mathbb{E}_{\boldsymbol{x}}(p_j(s^*, s^*), \boldsymbol{x})) = \frac{1}{2}$ to get:

$$s^*_{\varnothing} = \min\{1, \frac{\gamma + (1 - \gamma)(1 - e)}{2e\mathbb{E}_{\boldsymbol{x}}\big(f(c)\big)}\}$$

Since all players implement the same safety level $s^*_{\varnothing}$, the total expected probability of disaster is given by:

$$D_{\varnothing} = 1 - s^*_{\varnothing} = \max\{0, 1 - \frac{\gamma + (1 - \gamma)(1 - e)}{2e\mathbb{E}\big(f(c)\big)}\}$$

□

Proof of Proposition 9. Similar to the proof of Proposition 2, it proves convenient to continue with resulting performance level $k_i(x_i) \equiv x_i - s_i(x_i)$ employed by each player i.

The expected utility of a player with capability x_i is given by:

$$\begin{aligned} u_i(k_i(x_i)) = \gamma(x_i - k_i(x_i)) + (1-\gamma)\int_{-\infty}^{\infty} c - k(c) \quad \cdot f_{n-1}(c) \quad dc \\ + \int_{k^{-1}(k_i(x_i))}^{\infty} \Big(((1-e)(1-\gamma)-\gamma)(x_i - k_i(x_i)) + ((1-e)\gamma - (1-\gamma))(c-k(c))\Big) \\ \cdot f_{n-1}(c) \quad dc \end{aligned}$$

Taking the FOC with respect to $k_i(x_i)$ yields:

$$\begin{aligned} \frac{\partial u_i(k_i(x_i))}{\partial k_i(x_i)} &= -\gamma + (\gamma - (1-e)(1-\gamma)) \\ &\quad + \big(((1-e)(1-\gamma)-\gamma)F(x_i) + e(x_i - k_i(x_i))f(x_i)\frac{1}{k_i'(x_i)}\big) = 0 \\ \Rightarrow \quad e(x_i - k(x_i))f(x_i)\frac{1}{k'(x_i)} &= \underbrace{(1-e)(1-\gamma) + (\gamma - (1-e)(1-\gamma))F(x_i)}_{\Omega(x_i)} \\ k'(x) + \frac{e\cdot f(x_i)}{\Omega(x_i)}k(x_i) &= \frac{e\cdot f(x_i)}{\Omega(x_i)}x_i \end{aligned}$$

Following the same ODE and integration by parts steps in Proposition 2 gives:

$$k(x_i) = \Omega(x_i)^{-\frac{e}{\gamma-(1-e)(1-\gamma)}} \cdot \big(\Omega(x_i)^{\frac{e}{\gamma-(1-e)(1-\gamma)}} \cdot x_i - \int_{\underline{x}}^{x_i} \Omega(c)^{\frac{e}{\gamma-(1-e)(1-\gamma)}} dc + lim_{c\to\underline{x}}(k(c)-c)\cdot\Omega(c)^{\frac{e}{\gamma-(1-e)(1-\gamma)}}\mathrm{c}$$

$$s(x_i) = \frac{\int_{\underline{x}}^{x_i}\Omega(c)^{\frac{e}{\gamma-(1-e)(1-\gamma)}}dc}{\Omega(x_i)^{\frac{e}{\gamma-(1-e)(1-\gamma)}}} + lim_{c\to\underline{x}}s(c)\cdot\left(\frac{\Omega(c)}{\Omega(x_i)}\right)^{\frac{e}{\gamma-(1-e)(1-\gamma)}} + \text{const}$$

Note that we cannot choose $\underline{x} = -\infty$, because we must have that both sides of the ODE hold for all $x_i \in [\underline{x}, \mu]$. If not $lim_{c \to \underline{x}} s(c) \cdot \Omega(c)^{\frac{e}{\gamma-(1-e)(1-\gamma)}} \neq 0$ and it fails to hold. We select $\underline{x}$ as the minimum value of x such that the ODE obtains for all $x_i \in [\underline{x}, \mu]$.

Acknowledging that $s(\underline{x}) = 0$ and $s(x_i)$ is bounded below by 0 yields the final result:

$$s^*_{private}(x_i) = \min\{1, \frac{\int \Omega^{\frac{e}{\gamma-(1-e)(1-\gamma)}} dx_i}{\Omega^{\frac{e}{\gamma-(1-e)(1-\gamma)}}}\}$$

To check that the local optimum is a maximum, we evaluate the second order condition. Denoting the probability that a player with performance k_i wins as $P(k_i)$ and collecting terms, we have:

$$\frac{\partial^2 u_i(k_i)}{\partial k_i^2} = \frac{\partial^2 P(k_i)}{\partial k_i^2}(x_i - k_i)e - \frac{\partial P(k_i)}{\partial k_i}(2\gamma + 2e - e\gamma - 1)$$

By symmetry of strategies, we have that $P(k_i) = F(x_i)$. The second order condition then becomes

$$s^*(x_i)\frac{f'(x_i)}{f(x_i)} \leqslant \frac{1}{e}(2\gamma + 2e - e\gamma - 1)$$

□

Proof of Corollary 9.1. To solve for $\underline{x}$, we conjecture that $s^*(x_i)$ takes the form given in Proposition 9

$$s^*(x_i, \underline{x}) = \frac{\int_{\underline{x}}^{x_i} \Omega(c)^{\frac{e}{\gamma-(1-e)(1-\gamma)}} dc}{\Omega(x_i)^{\frac{e}{\gamma-(1-e)(1-\gamma)}}}$$

For a given $\underline{x}$ to check whether $s^*(x_i, \underline{x})$ holds, we turn to the second order conditions. Denote the function

$$I(s^*(x), \underline{x}) \equiv s^*(x)\frac{f'(x)}{f(x)} - \frac{1}{e}(2\gamma + 2e - e\gamma - 1)$$

In order for $s^*(x_i, \underline{x})$ to be a local maximum for some $\underline{x}$, we must have that $I(s^*(x), \underline{x}) \leqslant 0 \forall x \in [\underline{x}, \mu]$. Our optimization problem is then

$$argmin_{\underline{x} \in supp(C)} \underbrace{max_{x \in [\underline{x}, \mu]} I(s^*(x), \underline{x})}_{M(x, \underline{x})} \quad s.t. \quad max_{x \in [\underline{x}, \mu]} I(s^*(x), \underline{x}) \leqslant 0$$

We know that $M(x, \underline{x})$ is continuous and differentiable in $\underline{x}$ [Clarke, 1975]. Taking the derivative, we have

$$\frac{\partial}{\partial \underline{x}} max_{x \in [\underline{x}, \mu]} I(s^*(x), \underline{x}) = \underbrace{M_{\underline{x}}(x, \underline{x})}_{\geqslant 0} \underbrace{\left[-\frac{f'(x)}{f(x)} \left(\frac{\Omega(\underline{x})}{\Omega(x)} \right)^{\frac{e}{\gamma-(1-e)(1-\gamma)}} \right]}_{\leqslant 0} \leqslant 0$$

So M is weakly decreasing in $\underline{x}$. Consider $\gamma > 0, e > 0$. We have two cases. For $\gamma \geqslant \frac{1-2e}{2-e}$,

$M(x,\mu) = -\frac{1}{e}(2\gamma + 2e - e\gamma - 1) \leqslant 0$. Therefore, by the intermediate value theorem, we either have $\underline{x} = -\infty$ or $M(x,\underline{x})$ crosses 0 at a closed interval. We choose the lower bound of this integral to be the unique $\underline{x}$. Now consider $\gamma < \frac{1-2e}{2-e}$. Using the symmetry of $f(x)$ to note that $f'(x)$ becomes nonpositive for $x > \mathbb{E}[c] = \frac{\mu}{2}$. In this case, we can apply the same logic above to obtain $\underline{x}$. If the SOC are never satisfied on $supp(C)$, then $s^*(x) = 1 \forall x$.

□

Proof of Corollary 9.2. We know that when $x_i < \underline{x}$, the FOC do not hold and a corner solution obtains. Consider $s^*(x_i) = 0$. In this case $u_i(s_i) = 0$. In this case, a player has a unilateral incentive to deviate to $s^*(x_I) > 0$, so this is not an equilibrium. Therefore, $s^*(x_i) = 1$. □

Appendix B: Generalizing the effects of type distribution on risk

Though the main body of our text continues to set $G(x_i) \sim Uniform(0, \mu)$, our proofs use a more general type distribution, allowing us to compare how arbitrary player type distributions affect the overall level of disaster risk, generalizing the results for μ in section 7. We present the following two theorems.

Proposition 10. *Consider two non-noise adjusted random variables X_1, X_2 describing the distribution of player types with distribution functions $X_1 \sim G_1, X_2 \sim G_2$. Then expected disaster risk is lower under X_1 than X_2 if $X_1 \leqslant_{FOSD} X_2$.*

Proof of Proposition 10. Let $C_k = X_k + V_k - V_k \sim F(C_k)$ for $k = 1, 2$, where $V_k, V_k \sim_{i.i.d.} Logistic(0, \frac{1}{m})$ as in the main text. From Shaked and Shanthikumar [2007], we know that FOSD ordering is preserved under convolutions. Next we use the result that $X_1 \leqslant_{FOSD} X_2 \Leftrightarrow \mathbb{E}[\phi(X_1)] \leqslant \mathbb{E}[\phi(X_2)]$ for any increasing function $\phi(.)$. We therefore show that $s^*_{\varnothing}(X_k)$ is a weakly increasing function of X. Assuming that $X, f(C)$ are continuous and taking the derivative of Proposition 8, we have

$$\frac{\partial s^*_{\varnothing}}{\partial x} = \frac{[-\gamma - (1-\gamma)(1-e)]\, 2e \left(\lim_{b\to\infty} f(b)^2 - \lim_{a\to-\infty} f(a)^2\right)}{4e^2 \left(\mathbb{E}_{\boldsymbol{x}}\big(f(c)\big)\right)^2} = 0$$

□

Proposition 11. *Consider the same random variables X_1, X_2 as in Proposition 8. Then if $X_1 \leqslant_{FOSD} X_2$ and $1 \geqslant s(x)\frac{ef(x)}{\Omega}$, expected disaster risk is lower under X_1 than X_2.*

Proof of Proposition 11. Let C_k be defined as in Proposition 10. Following the same steps, we show that $s^*_{private}(x)$ is an increasing function of x. Assuming that $X, F(C)$ are continuous and taking the derivative of Proposition 9, we have

$$\begin{aligned}\frac{\partial s^*_{private}(x)}{\partial x} &= \frac{\Omega^{\frac{2e}{\gamma-(1-e)(1-\gamma)}} - \int \Omega^{\frac{e}{\gamma-(1-e)(1-\gamma)}} dx \frac{\partial}{\partial x}\Omega^{\frac{e}{\gamma-(1-e)(1-\gamma)}}}{\Omega^{\frac{2e}{\gamma-(1-e)(1-\gamma)}}} \\ &= 1 - s(x)\frac{ef(x)}{\Omega}\end{aligned}$$

Therefore, expected disaster risk is lower under X_1 than X_2 if and only if

$$s(x)\frac{ef(x)}{\Omega} \leqslant 1$$

□

Appendix C: Additional Figures

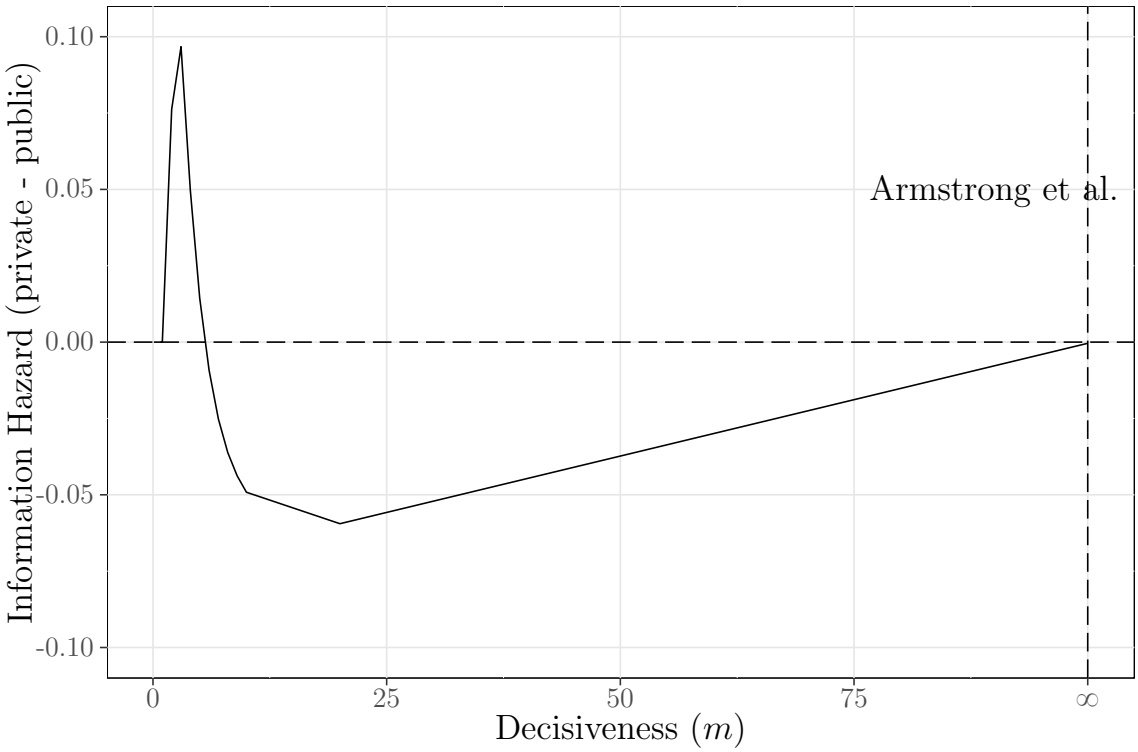

Figure 7: Information hazard (private info - public info)

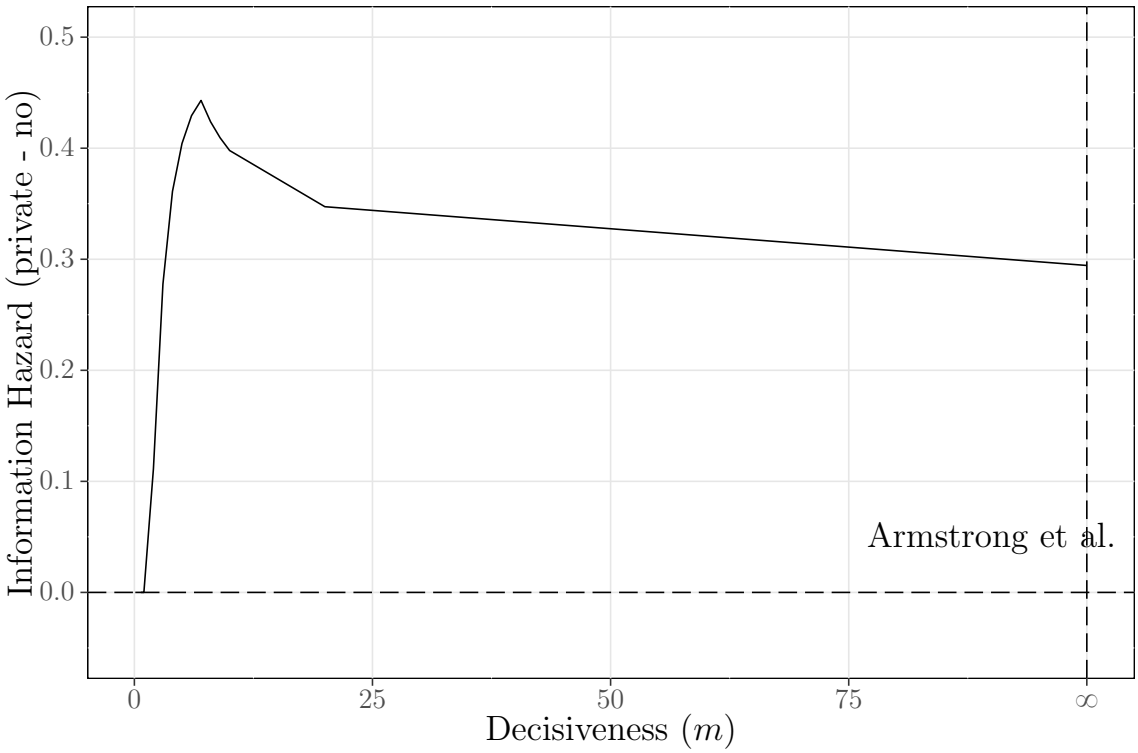

Figure 8: Information hazard (private info - no info)

References

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-worl Section: Technology.

Ajay Agrawal, Joshua Gans, and Avi Goldfarb. *The Economics of Artificial Intelligence: An Agenda*. University of Chicago Press, June 2019a. ISBN 978-0-226-61347-5. Google-Books-ID: 4GyVDwAAQBAJ.

Ajay Agrawal, John McHale, and Alexander Oettl. 5. Finding Needles in Haystacks: Artificial Intelligence and Recombinant Growth. In *The Economics of Artificial Intelligence: An Agenda*, pages 149–174. University of Chicago Press, June

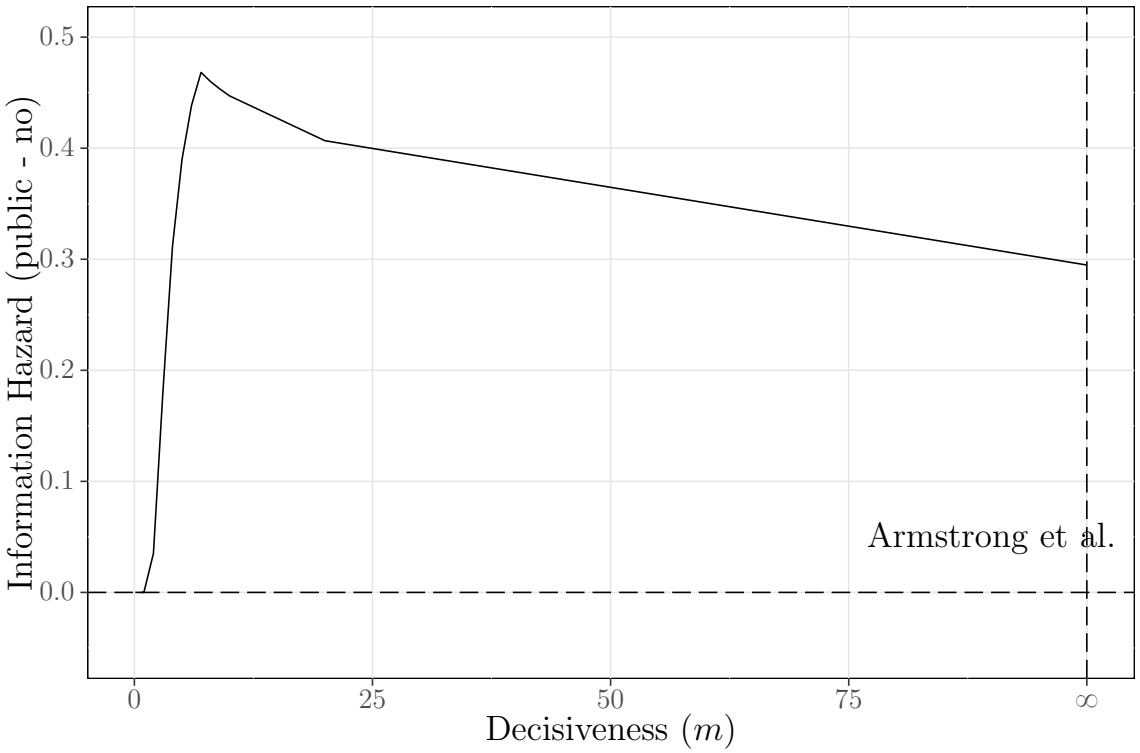

Figure 9: Information hazard (public info - no info)

2019b. ISBN 978-0-226-61347-5. doi: 10.7208/9780226613475-007. URL https://www.degruyter.com/document/doi/10.7208/9780226613475-007/html.

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. ISSN 0951-5666, 1435-5655. doi: 10.1007/s00146-015-0590-y. URL http://link.springer.com/10.1007/s00146-015-0590-y.

Leopold Aschenbrenner. Existential Risk and Growth. page 99, 2020.

Kyung Hwan Baik. Effort Levels in Contests with Two Asymmetric Players. *Southern Economic Journal*, 61(2):367, October 1994. ISSN 00384038. doi: 10.2307/1059984. URL https://www.jstor.org/stable/1059984?origin=crossref.

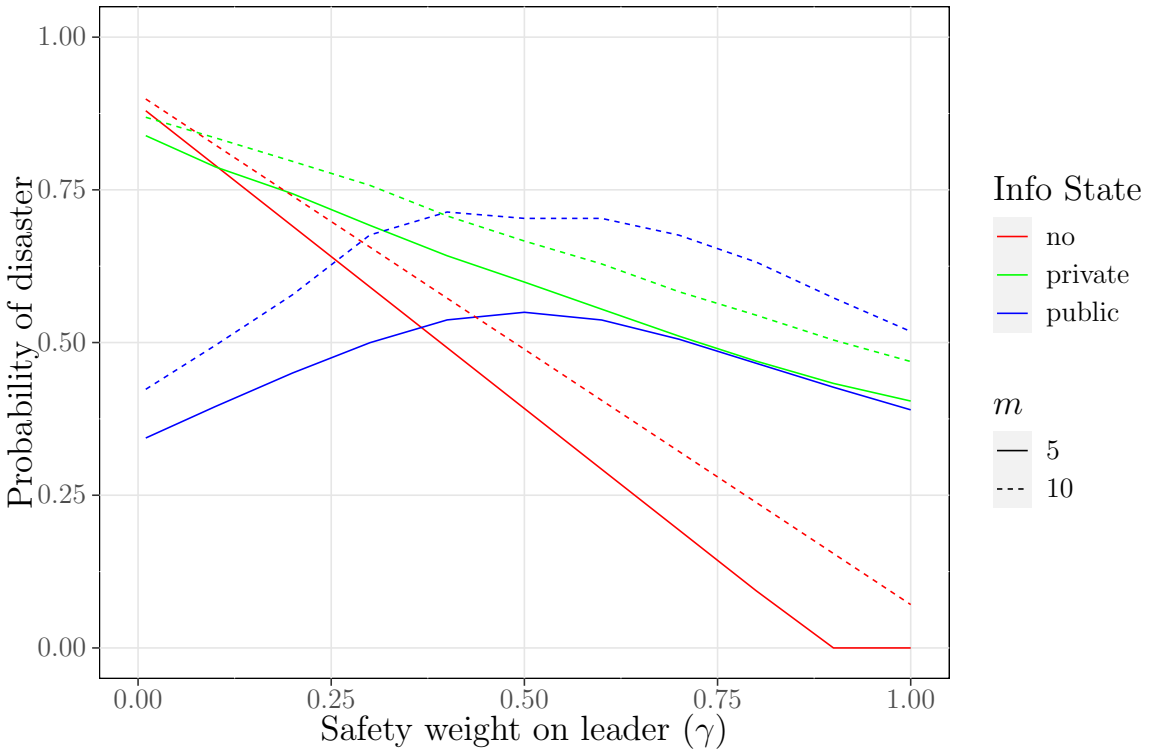

Figure 10: Varying safety contributions of the winner (full range of γ)

Muhammet A. Bas and Andrew J. Coe. A Dynamic Theory of Nuclear Proliferation and Preventive War. *International Organization*, 70(4):655–685, 2016. ISSN 0020-8183, 1531-5088. doi: 10.1017/S0020818316000230. URL https://www.cambridge.org/core/journals/international-organization/article/dynamic-the 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. ISSN 0899-8256. doi: 10.1016/S0899-8256(03)00027-7. URL https://www.sciencedirect.com/science/article/pii/S0899825603000277.

Martin Beraja, David Y. Yang, and Noam Yuchtman. Data-intensive Innovation and the

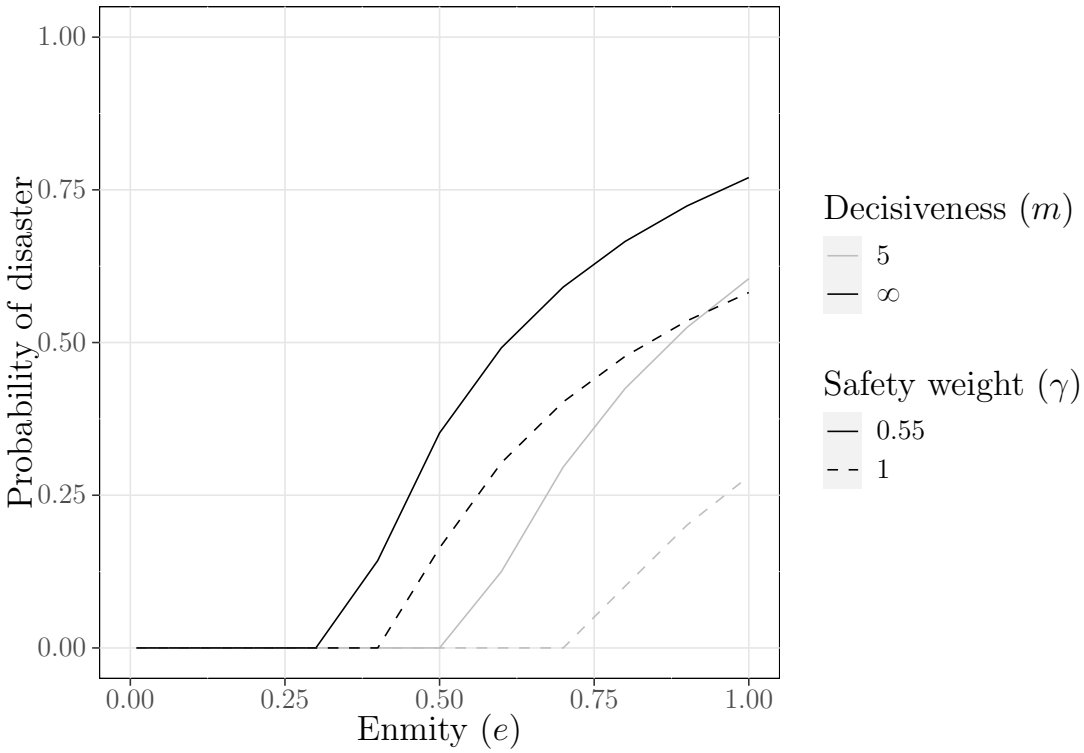

Figure 11: Effects of enmity under no information ($\mu = 0.72$)

State: Evidence from AI Firms in China. Working Paper 27723, National Bureau of Economic Research, August 2020. URL `https://www.nber.org/papers/w27723`. Series: Working Paper Series.

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

Nicholas Bloom. Are Ideas Getting Harder to Find? *THE AMERICAN ECONOMIC REVIEW*, 110(4):41, 2020.

Nick Bostrom. INFORMATION HAZARDS: A TYPOLOGY OF POTENTIAL HARMS FROM KNOWLEDGE. *Review of Contemporary Philosophy*, (10):44–79, 2011. ISSN

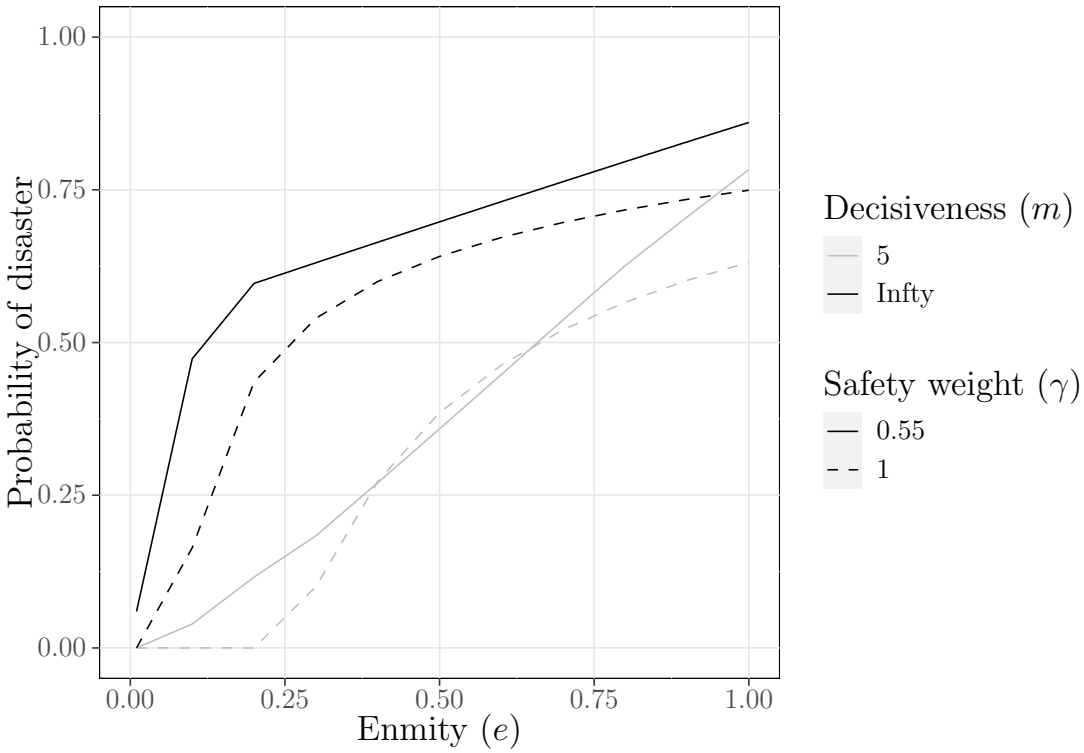

Figure 12: Effects of enmity under private information ($\mu = 0.72$)

1841-5261. URL https://www.ceeol.com/search/article-detail?id=44170. Publisher: Addleton Academic Publishers.

Nick Bostrom. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, 2014. ISBN 978-0-19-967811-2. Google-Books-ID: 7_H8AwAAQBAJ.

Nick Bostrom. Strategic Implications of Openness in AI Development. *Global Policy*, 8(2): 135–148, 2017. ISSN 1758-5899. doi: https://doi.org/10.1111/1758-5899.12403. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/1758-5899.12403. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1758-5899.12403.

Nick Bostrom. The Vulnerable World Hypothesis. *Global Policy*, 10(4): 455–476, 2019. ISSN 1758-5899. doi: 10.1111/1758-5899.12718. URL

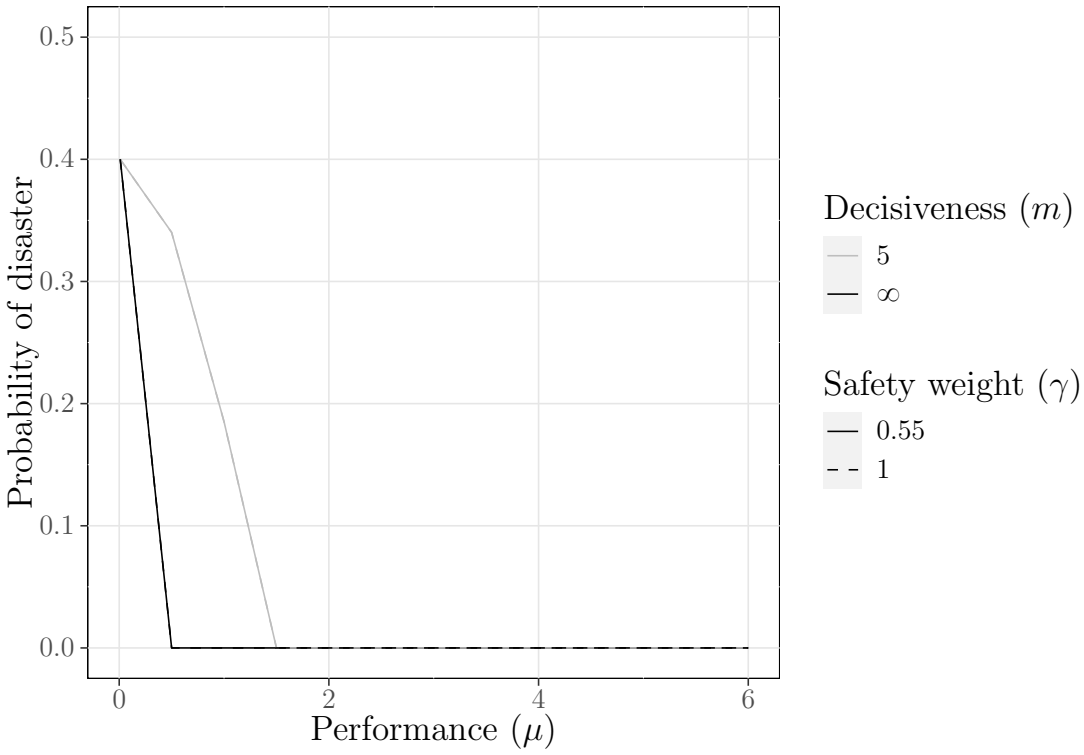

Figure 13: Effects of performance under no information.
We vary enmity so that $D_{\varnothing}(\mu = 0.01) = 0.4$. We have $e(m = 5, \gamma = 0.55) = 0.6897, e(m = 5, \gamma = 1) = 1, e(m = \infty, \gamma = 0.55) = 0.2249, e(m = \infty, \gamma = 1) = 0.2502$.

https://onlinelibrary.wiley.com/doi/abs/10.1111/1758-5899.12718. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1758-5899.12718.

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

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. ISSN 0899-8256. doi: 10.1006/game.1998.0709. URL https://www.sciencedirect.com/science/article/pii/S0899825698907096.

Frank H Clarke. Generalized gradients and applications. *Transactions of the American*

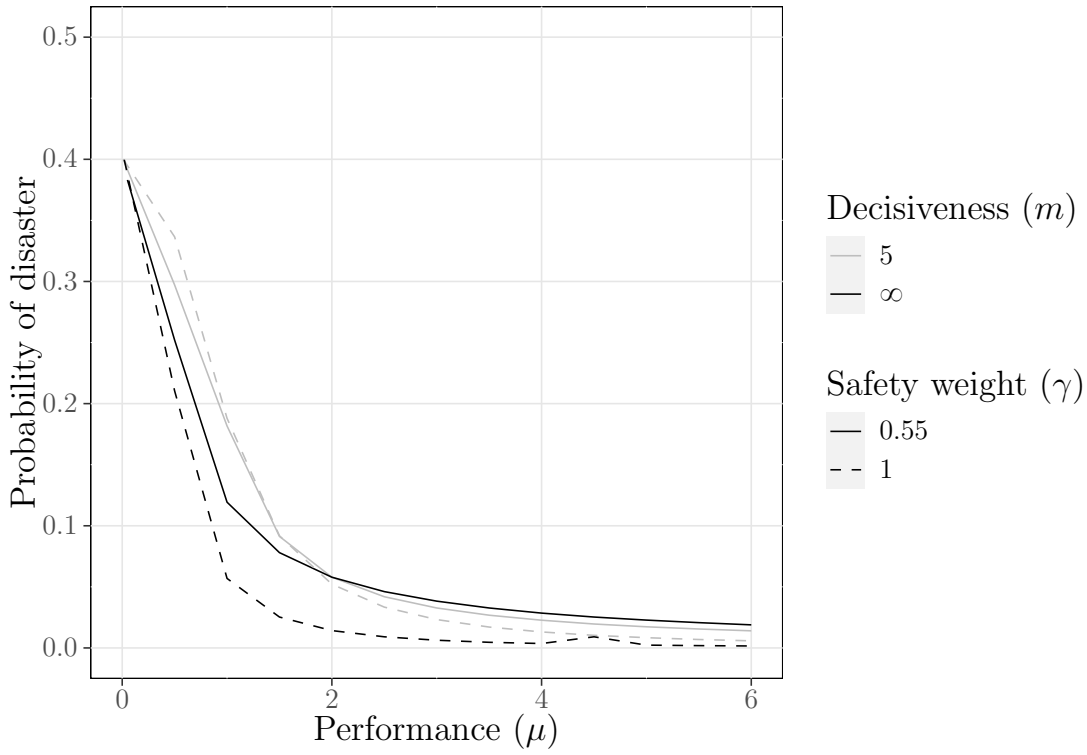

Figure 14: Effects of performance under private information
We vary enmity so that $D_{private}(\mu = 0.01) = 0.4$. We have $e(m = 5, \gamma = 0.55) = 0.339, e(m = 5, \gamma = 1) = 0.0.4035, e(m = \infty, \gamma = 0.55) = 0.0142, e(m = \infty, \gamma = 1) = 0.0881$.

Mathematical Society, 205:16, 1975.

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`.

Alexandre Debs and Nuno P. Monteiro. Known Unknowns: Power Shifts, Uncertainty, and War. *International Organization*, 68(1):1–31, January 2014. ISSN 0020-8183, 1531-5088. doi: 10.1017/S0020818313000192. URL `https://www.cambridge.org/core/product/identifier/S0020818313000192/type/journal`$_{a}rticle$.

E. Einy, O. Haimanko, D. Moreno, A. Sela, and B. Shitovitz. Equilibrium existence in Tullock contests with incomplete information. *Journal of Mathematical Economics*, 61:

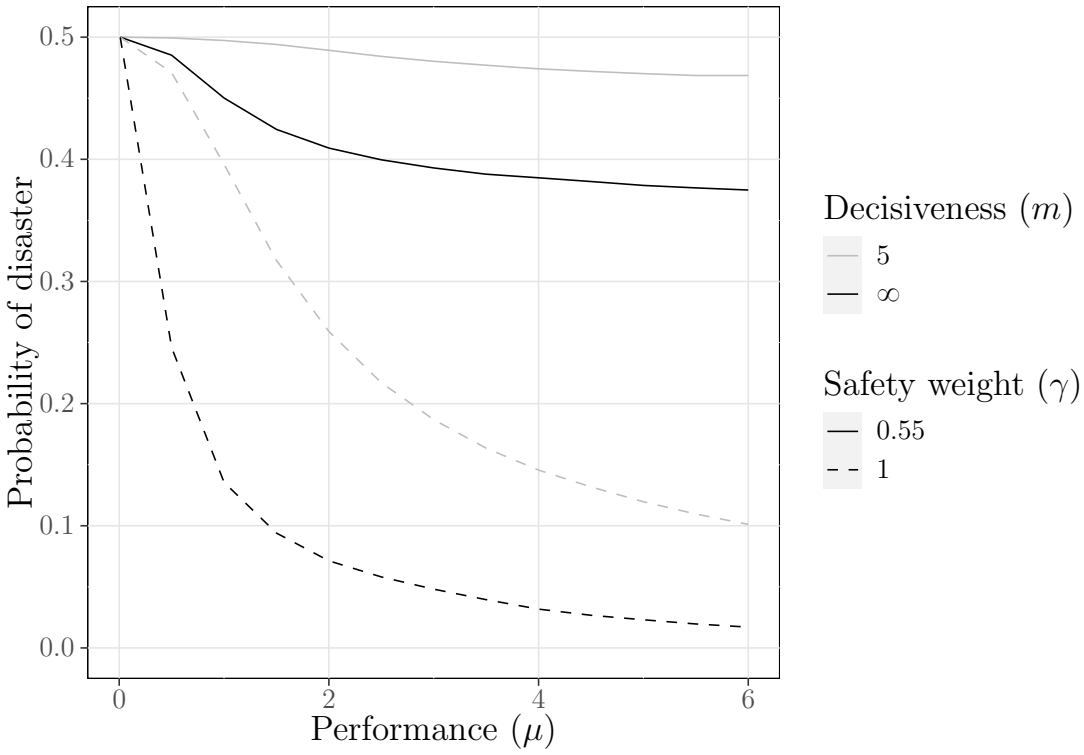

Figure 15: Effects of performance under public information
We vary enmity so that $D_{public}(\mu = 0.01) = 0.5$. We have $e(m = 5, \gamma = 0.55) = e(m = 5, \gamma = 1) = 0.8, e(m = \infty, \gamma = 0.55) = e(m = \infty, \gamma = 1) = 0.2$.

241–245, December 2015. ISSN 03044068. doi: 10.1016/j.jmateco.2015.10.004. URL https://linkinghub.elsevier.com/retrieve/pii/S0304406815001226.

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

Christian Ewerhart and Federico Quartieri. Unique equilibrium in contests with incomplete information. *Economic Theory*, 70(1):243–271, July 2020. ISSN 1432-0479. doi: 10.1007/s00199-019-01209-4. URL https://doi.org/10.1007/s00199-019-01209-4.

Sebastian Farquhar, John Halstead, Owen Cotton-Barratt, Stefan Schubert, Haydn Belfield,

and Andrew Snyder-Beattie. Existential Risk: Diplomacy and Governance, 2017. URL https://www.fhi.ox.ac.uk/wp-content/uploads/Existential-Risks-2017-01-23.pdf.

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.

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.

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. ISSN 0305-0629. doi: 10.1080/03050629508434870. URL https://doi.org/10.1080/03050629508434870. Publisher: Routledge _eprint: https://doi.org/10.1080/03050629508434870.

Martin Grossmann. Uncertain contest success function. *European Journal of Political Economy*, 33:134–148, March 2014. ISSN 0176-2680. doi: 10.1016/j.ejpoleco.2013.11.004. URL https://www.sciencedirect.com/science/article/pii/S0176268013000943.

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

ber 2000. ISSN 0022-3816. doi: 10.1111/0022-3816.00052. URL https://www.journals.uchicago.edu/doi/abs/10.1111/0022-3816.00052. Publisher: The University of Chicago Press.

Danny Hernandez. AI and Compute, May 2018. URL https://openai.com/blog/ai-and-compute/.

Jack Hirshleifer. Chapter 7 Theorizing about conflict. In *Handbook of Defense Economics*, volume 1, pages 165–189. Elsevier, January 1995. doi: 10.1016/S1574-0013(05)80009-2. URL https://www.sciencedirect.com/science/article/pii/S1574001305800092.

Robert Hogg, Joseph McKean, and Allen Craig. *Introduction to Mathematical Statistics*. Pearson, Boston, 7th edition edition, January 2012. ISBN 978-0-321-79543-4.

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

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. ISSN 0167-7187. doi: 10.1016/j.ijindorg.2012.06.007. URL https://www.sciencedirect.com/science/article/pii/S0167718712000811.

Charles I. Jones. R & D-Based Models of Economic Growth. *Journal of Political Economy*, 103(4):759–784, August 1995. ISSN 0022-3808. doi: 10.1086/262002. URL https://www.journals.uchicago.edu/doi/abs/10.1086/262002. Publisher: The University of Chicago Press.

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 McFadden. The measurement of urban travel demand. *Journal of Public Economics*, 3(4):303–328, November 1974. ISSN 0047-2727. doi: 10.1016/0047-2727(74)90003-6. URL https://www.sciencedirect.com/science/article/pii/0047272774900036.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys*, 54(6):1–35, July 2021. ISSN 0360-0300, 1557-7341. doi: 10.1145/3457607. URL https://dl.acm.org/doi/10.1145/3457607.

Adam Meirowitz. Strategic Uncertainty as a Cause of War. *Quarterly Journal of Political Science*, 3(4):327–352, December 2008. ISSN 15540634. doi: 10.1561/100.00008018. URL http://www.nowpublishers.com/article/Details/QJPS-8018.

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.

Wim Naude and Nicola Dimitri. The race for an artificial general intelligence: implications

for public policy. *AI & SOCIETY*, 35(2):367–379, June 2020. ISSN 1435-5655. doi: 10.1007/s00146-019-00887-x. URL https://doi.org/10.1007/s00146-019-00887-x.

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

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

Kristopher W. Ramsay. Information, Uncertainty, and War. *Annual Review of Political Science*, 20(1):505–527, May 2017. ISSN 1094-2939, 1545-1577. doi: 10.1146/annurev-polisci-051215-022729. URL https://www.annualreviews.org/doi/10.1146/annurev-polisci-051215-022729.

William Reed. Information, Power, and War. *American Political Science Review*, 97(4):633–641, November 2003. ISSN 0003-0554, 1537-5943. doi: 10.1017/S0003055403000923. URL https://www.cambridge.org/core/product/identifier/S0003055403000923/type/journal$_a$*rticle*.

Dmitry Ryvkin and Mikhail Drugov. The shape of luck and competition in winner-take-all tournaments. *Theoretical Economics*, 15(4):1587–1626, 2020. ISSN 1555-7561. doi: 10.3982/TE3824. URL https://onlinelibrary.wiley.com/doi/abs/10.3982/TE3824. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/TE3824.

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

Moshe Shaked and J. George Shanthikumar. *Stochastic Orders*. Springer Science & Business Media, April 2007. ISBN 978-0-387-34675-5. Google-Books-ID: rPiToBK2rwwC.

Stergios Skaperdas. Contest success functions. page 8.

Stergios Skaperdas. On the formation of alliances in conflict and contests. *Public Choice*, 96(1):25–42, July 1998. ISSN 1573-7101. doi: 10.1023/A:1004912124496. URL https://doi.org/10.1023/A:1004912124496.

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.

Neil C. Thompson, Kristjan Greenewald, Keeheon Lee, and Gabriel F. Manso. The Computational Limits of Deep Learning. *arXiv:2007.05558 [cs, stat]*, July 2020. URL http://arxiv.org/abs/2007.05558. arXiv: 2007.05558.

Robert Trager, Eoghan Stafford, Allan Dafoe, and Nicholas Emery. Strategic Implications of AI Safety Research, February 2021.

Gordon Tullock. *The Rent-seeking Society*. Liberty Fund, 2005. ISBN 978-0-86597-524-8. Google-Books-ID: w2_uAAAAMAAJ.

Robert Wiblin. Luisa Rodriguez on why global catastrophes seem unlikely to kill us all. URL https://80000hours.org/podcast/episodes/luisa-rodriguez-why-global-catastrophes-seem-u

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. ISSN 1540-5907. doi: 10.1111/j.1540-5907.2009.00388.x. URL

https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-5907.2009.00388.x.

_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1540-5907.2009.00388.x.

Eliezer Yudkowsky. Intelligence Explosion Microeconomics. page 96, 2013.

Susah Zhang and Mona Diab. Democratizing access to large-scale language models with OPT-175B. URL https://ai.facebook.com/blog/democratizing-access-to-large-scale-language-models-with-