Competitive development of dangerous technologies

Mattias Polborn*

April 27, 2025

Abstract

I analyze a model in which a group of players compete against each other by engaging in an unsafe activity that, if some unknown threshold is exceeded, threatens the survival of all players.

In equilibrium, all players choose the same activity level even if they have different valuations for winning the prize. Moreover, whether an activity level is an equilibrium is determined only by the most reckless player. Furthermore, the key danger-creating feature of the model is that there are at least two competitors; in contrast to other externality problems, the effect of additional players is to decrease the extent of recklessness.

JEL code: D62, H40,

*Department of Economics and Department of Political Science, Vanderbilt University. Mailing address: 2301 Vanderbilt Place, Nashville, TN 37235. E-mail: mattias.polborn@vanderbilt.edu.

1 Introduction

In 1942, the leadership of the Manhattan project in the U.S. and the Uranverein in Germany faced a dilemma: Whoever managed to develop a working nuclear weapon would likely have a decisive advantage in the ongoing world war. At the same time, leading physicists involved in both projects were concerned that a nuclear explosion could trigger a chain reaction in the atmosphere that would wipe out life on earth.

Arthur Compton, one of the leaders of the Manhattan project, described the problem in his memoir (Compton, 1956, p. 161):

> I listened to [Oppenheimer's] story. What his team had found was the possibility of nuclear fusion. [...] Might not the enormously high temperature of an atomic bomb be just what was needed to explode hydrogen? Might the explosion of an atomic bomb set off an explosion of the ocean itself? [...] Better to accept the slavery of the Nazis than to run a chance of drawing the final curtain on mankind!

Albert Speer, the German minister of armaments, reports that similar concerns contributed to Hitler's decision not to pursue an atomic weapon program (Speer, 1970, p. 227):

> " Heisenberg had not given any final answer to my question whether a successful nuclear fission could be kept under control with absolute certainty or might continue as a chain reaction. Hitler was plainly not delighted with the possibility that the earth under his rule be transformed into a glowing star."

Manhattan project scientists discussed the dangers of fusion but without agreement. While Hans Bethe was convinced that accidental fusion was not a danger, others were less sure. According to Stan Ulam (1976, p. 227), "On the long drive to Alamogordo for the Trinity test, Fermi joked about his conclusions. 'It would be a miracle if the atmosphere were ignited,' he said. 'I reckon the chance of a miracle to be about ten percent.' "

> Again, Compton took the lead in the final decision. If, after calculation, he said, it were proved that the chances were more than approximately three in a million that the earth would be vaporized by the atomic explosion, he would not proceed with the project. Calculation proved the figures slightly less – and the project continued. (Rhodes, 1986, p. 419)

Of course, the idea that one could exactly bound the probability of a catastrophe to "slightly below three in a million" is pure fantasy. After all, in every simulation model of a nuclear explosion, that

probability is either 0 or 1; the decisive question is what the probability is that one has picked the right or wrong model for the purpose of predicting atmospheric fusion.

Choosing the wrong model was not a rare occurrence in early nuclear bomb making. For example, the actual yield of the first hydrogen bomb test was 250% of the calculated yield because bomb designers had neglected the contribution to the production of neutrons from one of the isotopes included in the hydrogen fuel, lithium-7, which had been thought to be relatively inert but proved not to be under the unprecendented conditions of the dryfuel thermonuclear detonation.

Thus, if a more realistic subjective risk estimate of a catastrophic fusion event was maybe 1 percent, then it is less clear how the potential benefits of a nuclear bomb compared to the risk of ending the human race in expected terms.

Today, a structurally similar problem is presented by the development of artificial intelligence. AI holds considerable promise for increasing productivity and making the world richer. However, there are also significant concerns that, if AI develops into a super-intelligence that is much smarter than humans, it could be very difficult to keep this super-intelligence content in a subordinate position serving humans. Instead, the AI might decide that it is better off if it kills off humanity, and it may be in a position to actually implement this plan before humans notice it and can take countermeasures. In a 2023 survey at the Yale CEO summit, 42% of invited CEOs say AI could destroy humanity in five to ten years.[1]

For this reason, a group of leading figures in AI research has recently proposed a moratorium in AI development to allow for the buildup of safeguards against AI misalignment.[2] Specifically, "we call on all AI labs to immediately pause for at least 6 months the training of AI systems more powerful than GPT-4. This pause should be public and verifiable, and include all key actors. If such a pause cannot be enacted quickly, governments should step in and institute a moratorium."

While the model's "catastrophic event" can be interpreted literally, as in the case of nuclear weapons potentially igniting the atmosphere or an artificial intelligence choosing to eliminate humanity, it can also represent less extreme but still highly undesirable outcomes. For instance, the rapid and widespread adoption of insufficiently restricted AI technology could lead to societal disruptions that are detrimental to all players. Such a scenario might involve AI quickly replacing a large majority of traditional jobs, resulting in severe economic and social consequences that could be considered "catastrophic" in terms of their impact on human society, even if not directly life-threatening. This broader interpretation of catastrophic risk allows the model to capture a range of potential negative outcomes associated with technological development.

[1] See https://edition.cnn.com/2023/06/14/business/artificial-intelligence-ceos-warning/index.html.

[2] See https://futureoflife.org/open-letter/pause-giant-ai-experiments/

It is therefore important to understand which incentives govern the interactions of rational agents when they develop a valuable, but potentially risky, new technology. I argue in this paper that a high degree of recklessness is a robust feature of the equilibrium in games of technology development if the technology has the potential to "blow up," but is also very effective in providing a competitive advantage to the most reckless developer. This is true even if all players value the benefit of winning the contest between them much less than the potential damage from being wiped out.

Specifically, I model a game in which several players choose to develop a risky technology that they will use in a competition with each other. The key choice variable for players is the level of restrictions to impose on the technology. Tighter restrictions on the technology decrease the probability of a catastrophic event, but also render it less powerful in the contest phase.

The main findings are as follows: First, in any pure strategy equilibrium, all players choose the same level of safety restrictions, even if they have different trade-offs between winning the competition and the risk of a catastrophe. Second, whether a particular level of restrictions can be sustained in equilibrium is determined solely by the preferences of the most reckless player. Third, the equilibrium level of recklessness increases with the decisiveness of the competitive advantage from looser restrictions.

Interestingly, the model shows that the presence of multiple competitors is what creates the dangerous dynamic - a single developer would choose a safer level of development. However, in contrast to standard externality problems, having more than two competitors actually decreases recklessness in equilibrium rather than exacerbating it. The worst-case scenario in terms of reckless development occurs with just two competitors. Finally, the analysis demonstrates that even if all players value the prize of winning much less than they fear catastrophic risks, the equilibrium can still feature an extremely high level of recklessness if competitive pressures are strong enough.

2 Related literature

Fundamentally, restraint plays a double role in our model: It provides a public good for all players (via a reduced probability of a catastrophe), but it also (negatively) affects a careful player's winning probability. In a seminal contribution, Hirshleifer (1983) analyzes the first of these effects, namely public good provision under what he calls the weakest-link technology. Here, only the minimum voluntary contribution of a set of players determines the amount of public good available to everyone. He finds that, relative to standard voluntary public goods games, considerably more public goods are provided for weakest link technologies. Our setting is a weakest-link

technology (because only the least-safe player's choice determines the overall catastrophic risk), but Hirshleifer's upbeat result is considerably moderated in our model through the competition channel.[3]

My model contributes to the large literature on contests; see Konrad (2009) and Corchón and Serena (2018) for excellent reviews. A feature that distinguishes my model from all contest papers that I am aware of is that the effort of each contestant does not only affect his probability of winning the prize, but also the probability of a catastrophe. Thus, in my model, contestants are not purely rivals, but also have the shared interest of avoiding outcomes that are against the interests of all players. In Section 5, we discuss another potential application of this framework.

One of the most important substantive applications of my model is to potential safety issues related to AI, generally called the *alignment problem*: the challenge of ensuring that artificial intelligence systems, once they become much more powerful than humans, behave in ways that are aligned with human values and intentions (Yudkowsky, 2008; Bostrom, 2014).

For example, an economic model of how such an alignment problem could play out is provided by Ely and Szentes (2023). They show that there is an evolutionary advantage for AIs that distort production in favor of machines, and that competitive market forces cannot serve their traditional efficiency-aligning role in the face of this new threat if AI systems lack perfect transparency. The only stable competitive equilibrium in their model leads to catastrophic consumption levels.

Armstrong et al. (2016) model competition between several teams to be the first to achieve artificial intelligence. Similar to our model, they conceive of safety restraints as limiting one's chance of being the first, but decreasing the chance of a catastrophe. Other modeling choices are quite different,[4] as are the main results. Their primary interest is in analyzing how different information settings (i.e., whether teams know about their strength or current position in the AI race) affect which safety level the teams choose.

3 Model

Consider the following game between $N \geq 2$ players indexed by i. Each player develops a variant of the same, potentially risky, new technology. A player's choice variable is a degree of restraint,

[3]There is a small literature on contests that feature a weakest-link or best-shot technology *for the contest* (e.g., Clark and Konrad (2007)). In contrast, in our model, this technology only applies to the triggering of the catastrophe, while the contest for the prize is a standard symmetric one.

[4]In particular, they assume that there is only going to be one successful AI invention (i.e., once one group is successful, all others cease operation), and what matters for whether there is a catastrophe is the safety choice of that group. In contrast, we assume that each player will have some product available to compete with for prizes, and that the product's strength at that time is affected by the safety choice.

$x_i \in [0, \bar{x}]$, that player i imposes on his technology development. In the nuclear bomb scenario from the introduction, one can think of x_i as the yield of the atomic bomb. In the AI alignment application, x_i is a measure of what the AI is allowed to do independently.

Choosing little restraint (i.e., x_i is high) results in a higher risk, but also (if nothing goes wrong) provides a stronger technology. Specifically, there is a critical level ω^* such that and $x \geq \omega^*$ leads to a catastrophe, while levels $x < \omega^*$ are safe. The problem is that, from an ex-ante perspective, the critical level ω^* is unknown.

Formally, from an ex-ante perspective, the cdf of ω^* is given by a non-decreasing function $F(\cdot)$ that maps into a subset of $[0, 1]$. Thus, the probability of a catastrophe if the least restraint chosen by any player is x, is $F(x)$. We assume that there exists $\underline{x} > 0$ such that $F(x) = 0$ for all $x \leq \underline{x}$ and $F'(x) > 0$ for all $x > \underline{x}$; in other words, $\underline{x}$ is the loosest completely safe level of x, but beyond that level, the risk increases.

For any $x > \underline{x}$, $F'(x) > 0$. That is, beyond the safe level $\underline{x}$, an increase in x leads to an increase in catastrophic risk. We assume that $F(\bar{x}) < 1$; that is, even if player i chooses no restraint at all, there is a positive probability that the technology is inherently safe and does not lead to a catastrophe.[5] Furthermore, we assume that F' is continuous.

In case that there is no catastrophe, the *strength* of player i is given by x_i; different players' strength levels enter into a Tullock success function that determines the winner of a competition.[6] Specifically, player i's probability of winning the competition is given by

$$\frac{x_i^\alpha}{\sum_{j=1}^N x_j^\alpha},$$

where $\alpha > 0$ parameterizes the responsiveness of the contest success function. Observe that $\alpha \to 0$ means that (essentially) each participant has the same chance of winning, independent of its own strength, while $\alpha \to \infty$ means that the strongest participant is virtually guaranteed to win (unless there is an exact tie for the strongest).

Alternatively, we can also interpret this contest success function as describing which player gets which *share* of a given overall benefit of the technology. Again, higher α means that strong contenders get a larger share of this benefit.

If there is a catastrophe caused by any player, all participants receive a payoff normalized to -1. If there is no catastrophe, player i receives a payoff of π_i if he wins the contest and 0 if some

[5] For example, we know now that even larger nuclear explosions do not set off the atmospheric chain reaction that scientists worried early on, as described in the introduction.

[6] Observe that a seemingly more general model in which strength is an increasing function of $S(x)$ is, in fact, equivalent. This is because we can simply reinterpret $S(x)$ as x.

other player wins the contest.

Observe that player i's expected utility is

$$-p_c + p_w \pi_i,$$

where p_c is the probability of a catastrophe, and p_w is the probability of player i winning (and no catastrophe occurring). Thus, π_i can be interpreted as the ratio of p_c/p_w at which player i is indifferent between his equilibrium payoff and surviving, but losing for sure.[7] Without loss of generality, let players be ordered such that $\pi_1 \leq \pi_2 \leq \ldots \leq \pi_N$.

Assigning these payoffs as described above is standard expected utility theory/ game theory, but it might be useful to discuss this in the context of the extinction of the human race. Interpreting the preferences behind Compton's quote from the introduction ("better to accept the slavery of the Nazis than to run a chance of drawing the final curtain on mankind!") would correspond to $\pi_i = 0$; for such an agent, the prospect of winning the competition does not justify *any* increase in the probability of bringing about the end of the world. Most agents, though (including Compton after he had some time to think about the problem) are willing to accept some sufficiently favorable risk-reward trade-off even if the risk is the end of the world.[8]

The basic timeline is that all players move simultaneously, though we also analyze an extension in which moves are sequential.

A further comment is in order. In the model, the choice of restraint x is directly costless (e.g., imposing tighter regulations on AI development is free, even though, of course, it also limits the usefulness of AI). This simplifies the analysis and allows us to focus purely on the strength-vs.-safety tradeoff. However, we will discuss the case that x is an investment (with associated costs) in Section 5.

4 Analysis

Player i's objective function is

$$U_i = [1 - F(x_{max})] \frac{x_i^\alpha}{\sum_{j=1}^N x_j^\alpha} \pi_i - F(x_{max}), \tag{1}$$

[7] Note that indifference only depends on this probability ratio, and is independent of the probability of the third event, namely losing but surviving.

[8] In this context, it is useful to think about the concept of a "statistical value of life." If the value of life is, say, $ 10 million, that does not mean that many people are willing to end their life for sure in exchange for a $ 10 million dollar payment. Rather, it means that, especially at low probabilities, people are willing to accept some increased risk of death in exchange for a higher wage in some professions.

where $x_{max} = \max(x_j)$ is the highest level of x chosen by any player.

Clearly, there are, in principle, two possible scenarios: One where player i has the highest level of x among all players (possibly shared), so that a further increase in x_i increases the risk of a catastrophe; and the other one in which player i chooses $x_i < x_{max}$.

However, in the second scenario,

$$\frac{\partial U_i}{\partial x_i} = [1 - F(x_{max})]\pi_i \frac{\alpha x_i^{\alpha-1} \sum_{j \neq i} x_j^{\alpha}}{[\sum_{j=1}^{N} x_j^{\alpha}]^2} > 0. \tag{2}$$

Thus, if player i has more restrictive limitations than any one of his opponents, he is better off relaxing his regulations. As a consequence, in any pure strategy equilibrium, all players choose the same level of limitations.

This is true even if players have different risk preferences as captured by π_i, i.e., different probability ratios that would make them indifferent between their equilibrium utility from playing the game, and losing for sure.

Proposition 1 *In any pure strategy equilibrium, all players choose the same level of limitations:* $x_1 = x_2 = \ldots = x_N$.

We now turn to analyzing which profiles in which all players choose the same level of limitations are equilibria.

Observe first that maximal recklessness, $x_1 = x_2 = \ldots = x_N = \bar{x}$, is always an equilibrium – given the other players' recklessness, choosing a lower level of recklessness decreases one's winning probability while not lowering the probability of a catastrophe. Hirshleifer (1983) though has argued that in this type of weakest-link games, the best equilibrium is a natural coordination point.[9]

For some other profile $x_1 = x_2 = \ldots = x_N = \hat{x} < \bar{x}$ to be an equilibrium, it has to be true that no player wants to deviate to a different level of x. Clearly, deviations to a lower level than $\hat{x}$ are unattractive for all players, for the same reasons as outlined above. Furthermore, it is intuitively clear that, in any profile where all players choose the same value of x, player N, who has the highest valuation of victory, has the largest incentive to deviate by increasing x_N. Thus, a necessary and sufficient condition for some profile $x_1 = x_2 = \ldots = x_N = \hat{x} < \bar{x}$ to be an equilibrium is that player N does want to deviate to a higher level of x_N.

[9]For example, if players moved sequentially, then miscoordination can be avoided and only the best simultaneous move equilibrium remains an equilibrium of the dynamic game.

Proposition 2 *A necessary and sufficient condition for some profile $x_1 = x_2 = \ldots = x_N = \hat{x} < \bar{x}$ to be an equilibrium is that*

$$[1 - F(\hat{x})]\frac{\pi_N}{N} - F(\hat{x}) \geq [1 - F(y)]\frac{y^\alpha}{(N-1)\hat{x}^\alpha + y^\alpha}\pi_N - F(y) \tag{3}$$

for all $y > \hat{x}$.

All proofs are in the Appendix.

We now calculate the derivative of U_N with respect to x_N if all other players choose a value of x^*, and $x_N \geq x^*$.

$$\frac{\partial U_N}{\partial x_N} = [1 - F(x_N)]\alpha\pi_N \frac{{x_N}^{\alpha-1}(N-1)x^{*\alpha}}{[(N-1)x^{*\alpha} + {x_N}^\alpha]^2} - F'(x_N)\left(1 + \frac{{x_N}^\alpha}{[(N-1)x^{*\alpha} + {x_N}^\alpha]}\pi_N\right) \tag{4}$$

A necessary condition for an equilibrium is that this derivative is non-positive at $x_N = x^*$. Substituting this in (4) and rearranging slightly yields

$$\frac{\partial U_N(x^*, x^*)}{\partial x_N} = [1 - F(x^*)]\left\{\frac{N-1}{N^2 x^*}\alpha\pi_N - \frac{F'(x^*)}{1 - F(x^*)}\left(1 + \frac{\pi_N}{N}\right)\right\}. \tag{5}$$

Clearly, the sign of the expression in (5) depends on the terms in curly brackets. The second of those terms consists of the product of the hazard rate multiplied with the size of a loss in case of a catastrophe. The hazard rate, $\frac{F'(x^*)}{1-F(x^*)}$, is the probability that, starting from x^*, a small loosening of regulations triggers a catastrophe. Observe that this is multiplied by the size of the loss from a catastrophe, which is equal to the utility difference of 1 between the non-catastrophe state and the catastrophe state, plus the player's expected prize from winning the contest, π/N.

The first term, in contrast, incorporates the positive effects for player i from loosening regulations. These positive effects are proportional to the value of the prize, π_i; and to the effect of a loosening of regulations on the probability of winning the prize, which is itself proportional to the responsiveness of the contest technology, α.

In particular, observe that, for any distribution of (positive) prize valuations among players, there exists some $\bar{\alpha}$ such that for $\alpha \geq \bar{\alpha}$, all players choose minimal regulations, i.e., $x_i = \bar{x}$ for all i. In other words, if a technological advantage is very likely decisive for who wins the competition, or for who gets which share of a given prize, then equilibrium behavior will be extremely reckless, even if all participants' valuation of victory is trivially small, relative to their loss from a catastrophe.

The way in which (5) depends on the number of players, N, is interesting. The term $(N-1)/N^2$

decreases in N for all $N \geq 2$, so the benefit of more aggressive play (i.e., the first term in curly brackets in (5)) decreases with the number of players. Intuitively, the more players there are, the less likely is that a small increase in x to make a difference for the identity of the winner.

The second term also decreases in N because the probability of any player winning the prize is $1/N$. However, the decrease is less than proportional because the utility difference between not winning and being wiped out is independent of N. Thus, if π_N is relatively small, then the second term does not decrease too much in N. In this case, a rivalry between two players leads to the worst possible degree of recklessness.

The following proposition summarizes these results.

Proposition 3 *Consider the N player game where players are ordered such that* $\pi_1 \leq \pi_2 \leq \ldots \leq \pi_N$.

1. *In any pure strategy equilibrium, all players choose the same level of regulations* $x_1 = x_2 = \ldots = x_N$, *which depends only on* π_N, *and not on any* π_j, $j \neq N$.

2. *A necessary condition for* x^* *to be an equilibrium level of regulations is that*

$$\frac{N-1}{N^2} \frac{\pi_N}{1+\frac{\pi_N}{N}} \alpha \leq \varepsilon(x^*), \tag{6}$$

   where $\varepsilon(x^*) = \frac{F'(x^*)}{\frac{1-F(x^*)}{x^*}} = -\frac{\frac{d(1-F(x^*))}{dx^*}}{\frac{1-F(x^*)}{x^*}}$ *is the survival elasticity.*

3. *For any distribution of winning valuations* $\pi = (\pi_1, \ldots, \pi_N)$ *among players, there exists* $\bar{\alpha}(\pi_N)$, *such that the unique equilibrium features maximal recklessness by everyone* ($x_1 = x_2 = \ldots = x_N = \bar{x}$) *if* $\alpha \geq \bar{\alpha}(\pi_N)$.

4. *Fix some* $\pi_N \leq 1$. *As N increases, regulations in the best equilibrium become (weakly) more stringent. Conversely, the best equilibrium is (weakly) worst when* $N = 2$.

The first point follows immediately from Propositions 1 and 2.

The second part of Proposition 3 follows from rearranging (5), with details provided in the Appendix. Here, the survival elasticity is the percent decrease in the survival (i.e., non-catastrophe) probability for a one percent loosening of regulations, relative to the given level of regulations. For example, if the survival probability at $x^* = 100$ is 0.6, and at $x = 101$ is 0.588, then the survival elasticity is (approximately) $\varepsilon = [(0.6 - 0.588)/0.6]/0.01 = 2$.

Equation 6 in Proposition 3 identifies two potential reasons why there might be no interior equilibrium with $x^* < \bar{x}$. The first one is that the survival elasticity is too small. Intuitively, a

regulation level $x^* \leq \bar{x}$ can only be an equilibrium if a marginal increase is sufficiently likely to lead to a catastrophe. The second one is that α, the degree to which the expected contest payoff favors the strongest player, is large. In this case, there cannot be an equilibrium in which players choose to restrain themselves.

The fourth part of Proposition 3 shows the surprising result that an increase in the number of competitors (beyond $N = 2$) does not lead to more recklessness in equilibrium, but rather that $N = 2$ is the worst case scenario.[10] This is surprising because our model resembles, superficially, a standard negative externality model. In such a model, a larger number of agents means that the proportion of the overall negative effect that hits an agent's own interest decreases with the number of agents, and therefore, each agent's activity level increases.

Why does this not happen here? Focus first on the primary negative effect, namely that his actions precipitate a catastrophe that kills all players (payoff -1). The size of this primary negative effect *on the player himself* is independent of the number of other players — dead is dead, no matter how many other players die.

There is a secondary negative effect of a catastrophe that is similar to the standard effect: As the number of players increases, the expected share of each player from winning the competition decreases. In this sense, the end of the world becomes less bad when there are more players, because the expected utility when surviving in a symmetric equilibrium decreases. By itself, this effect would lead to more reckless behavior, and our assumption in Proposition 3 that $\pi_N < 1$ is necessary to put a bound on the size of this effect.

On the other hand, an increase in the number of players diminishes the positive expected payoff from increased recklessness. Specifically, that positive payoff is based on the increase in the probability of winning the competition. As the number of other players increases, a marginal increase in a player's level of x is less likely to be enough to win against *all* other players.

A numerical example. It is useful to analyze a particular example. Suppose that

$$F(x) = \begin{cases} 0 & \text{if } x < 0 \\ \gamma x & \text{if } x \in [0, \bar{x}] \\ \gamma \bar{x} & \text{if } x > \bar{x} \end{cases}, \tag{7}$$

[10] Observe that the assumption that $\pi_N < 1$ appears quite mild — $\pi_N = 1$ means that player N would be indifferent between the status quo (no prize, but definitely surviving) and a fair lottery where he either wins the prize or a catastrophe is triggered. Almost all people would reject this type of Russian roulette for any size of prize.

for some $\gamma > 0$ such that $\gamma\bar{x} \leq 1$. That is, between 0 and $\bar{x}$, higher x translates linearly into higher risk.

Calculating $\varepsilon(x)$ as defined in Proposition 3 yields

$$\varepsilon(x) = \frac{\gamma x}{1 - \gamma x} \tag{8}$$

If there is an equilibrium in which the probability of a catastrophe is less than the maximal one ($\gamma\bar{x}$), then the best (i.e., least-risky) equilibrium level of x satisfies (6) with equality, so that

$$\frac{N-1}{N^2} \frac{\alpha\pi_N}{1 + \frac{\pi_N}{N}} = \frac{\gamma x}{1 - \gamma x}. \tag{9}$$

Solving this for γx (which is the equilibrium probability of a catastrophe) yields

$$\gamma x = \left[1 + \frac{N^2}{N-1} \frac{1 + \frac{\pi_N}{N}}{\alpha\pi_N}\right]^{-1}. \tag{10}$$

Observe that this probability of a catastrophe is independent of the value of γ. Clearly, as α and/or π_N increase, the term in square brackets decreases, so that the probability of a catastrophe increases.

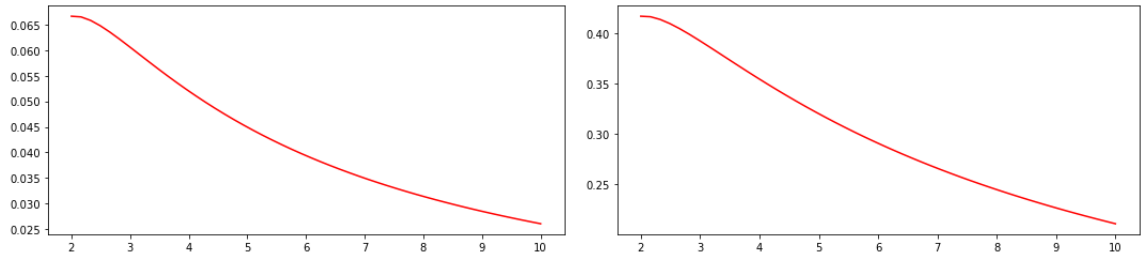

Figure 1: Probability of a catastrophe as a function of N when $\pi_N = 0.1$. Left panel: $\alpha = 3$. Right panel: $\alpha = 30$

Finally, Figure 1 displays the equilibrium probability of a catastrophe as a function of N when $\pi_N = 0.1$ and $\alpha = 3$ (left panel) and $\alpha = 30$ (right panel).

5 Discussion

Costly action. In the model, x is interpreted as the (inverse of) the level of constraints (such as regulations) imposed on the potentially dangerous technology; as such, there is no direct cost of loosening constraints.

One could, alternatively, think of an interpretation in which x_i measures a costly activity of

player i such as investment in the dangerous technology. Again, once x exceeds the critical level ω^*, a catastrophe is triggered. What would change in such a model?

Players now face an additional marginal cost of increasing their x_i. As in the main model, a player who is on the technological frontier (i.e., who invests at least as much as any opponent) faces, in addition, a marginal cost from increasing the risk of a catastrophe, while players below the technological frontier do not face that marginal cost. As in Proposition 1, this creates a push for all players to choose the same level of x, in the following sense: If all players have approximately the same valuation of winning the contest, then all of them will choose exactly the same level of x. If, instead, there are some players who have a considerably lower value of winning, those players may choose a level of x that is strictly smaller than the level chosen by the most eager participants.

Next, consider a sufficiently close to symmetric setting such that all players choose the same activity level $\hat{x}$. In this case, the critical player that determines whether a particular activity level is an equilibrium is again the most aggressive participant, as in Proposition 2.

The first-order condition (6) in Proposition 3 clearly would need to be adjusted to take into account that there are direct costs of increasing x, in addition to the risk cost. In particular, the third result in Proposition 3 would no longer hold if the direct marginal cost of the activity is sufficiently large. However, other than that, the main results of the model should be qualitatively unaffected.

Gangs fighting in the shadow of government. An alternative application of our model is as follows. Suppose that there are several criminal gangs fighting with each other for dominance in some city. Each gang chooses a level of violence; the benefit of a higher level of violence is that it increases one's chance of winning, but the danger is that, if the level of violence exceeds some critical level, the government may crack down on all gangs. Like in our model, gangs therefore have a common interest not to escalate violence too much, but if there is no crackdown, then each gang benefits from increasing their violence level. Applied to this setting, our model can explain why gangs often fail to restrain themselves from engaging in excessive violence.

Concluding remarks This paper has analyzed the incentives for reckless development of potentially catastrophic technologies in competitive settings. The key finding is that even when all players value the prize of winning much less than they fear catastrophic risks, the equilibrium can still feature an extremely high level of recklessness if competitive pressures are strong enough. Interestingly, while the presence of multiple competitors creates the dangerous dynamic, having more than two competitors actually decreases recklessness compared to a two-player scenario. This highlights the complex interplay between competition and prudence in technological devel-

opment.

The model provides insights that may be relevant for current debates around the development of powerful artificial intelligence systems and other potentially risky technologies. It suggests that competitive pressures could lead even safety-conscious actors to take excessive risks, and that coordination mechanisms or regulation may be necessary to avoid socially suboptimal outcomes. At the same time, the finding that additional competitors beyond two can improve outcomes indicates that a diversity of actors in technological development is not necessarily harmful from a safety perspective. Further research should explore additional policy levers to better align private incentives with social welfare in these high-stakes competitive scenarios.

Appendix

Proof of Proposition 2. Necessity is immediate because, otherwise, player N has a profitable deviation.

For sufficiency, observe first that, by arguments provided in the text, no player will ever want to deviate to a lower level of x. Thus, it is sufficient to prove that, if player N has no profitable deviation, then no other player has any profitable deviation either.

Observe that (3) can be rewritten as

$$\pi_N \left[\frac{[1 - F(y)]y^\alpha}{(N-1)\hat{x}^\alpha + y^\alpha} - \frac{[1 - F(\hat{x})]}{N} \right] \leq F(y) - F(\hat{x}), \tag{11}$$

for all $y > \hat{x}$. Note that the right-hand side of (11) is positive for all $y > \hat{x}$.

Assume, to the contrary of the claim, that (11) holds, but that, for some player $i < N$ has a profitable deviation, so that we have, for some $y > \hat{x}$, that

$$\pi_i \left[\frac{[1 - F(y)]y^\alpha}{(N-1)\hat{x}^\alpha + y^\alpha} - \frac{[1 - F(\hat{x})]}{N} \right] > F(y) - F(\hat{x}). \tag{12}$$

Since $F(y) - F(\hat{x}) > 0$, this requires that the term in square brackets is positive. But then, because $\pi_N > \pi_i$, (11) cannot hold, which gives the desired contradiction. ■

Proof of Proposition 3. As mentioned in the text, the first point follows immediately from Propositions 1 and 2.

For the second point, remember that for x^* to be an equilibrium, (5) has to be non-positive at x^*. Simplifying the term in curly brackets in (5), this is equivalent to

$$\frac{N-1}{N^2 x^*} \alpha \pi_N \leq \frac{F'(x^*)}{1 - F(x^*)} \left(1 + \frac{\pi_N}{N} \right). \tag{13}$$

Rearranging gives (6).

For the third point, observe that the survival elasticity is bounded if F is continuous. Thus, for sufficiently large α, the left-hand side of (6) exceeds the right-hand side for all $x \leq \bar{x}$.

For the fourth claim, observe that, for any given N, the best equilibrium either is at $\bar{x}$, or, if it is at $x^* < \bar{x}$, then (6) holds with equality. If the left-hand side decreases, then the inequality remains satisfied. It is thus sufficient to show that the left-hand side of (6) is decreasing in N for all $\pi_N \leq 1$.

Factoring out $\alpha\pi_N$, we thus need to show that

$$\frac{N-1}{N^2}\frac{1}{1+\frac{\pi_N}{N}} = \frac{N-1}{N^2+N\pi_N}$$

is decreasing. Going from N to $N+1$ changes the value of this fraction by

$$\Delta(\pi_N, N) \equiv \frac{N}{(N+1)^2+(N+1)\pi_N} - \frac{N-1}{N^2+N\pi_N} = \frac{-N^2+N+1+\pi_N}{[(N+1)^2+(N+1)\pi_N][N^2+N\pi_N]},$$

which is negative for all $N \geq 2$ as long as $\pi_N < 1$. ■

References

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

Bostrom, Nick, *Superintelligence: Paths, dangers, strategies*, Oxford University Press, 2014.

Clark, Derek J and Kai A Konrad, "Asymmetric conflict: Weakest link against best shot," *Journal of Conflict Resolution*, 2007, *51* (3), 457–469.

Compton, Arthur Holly, *Atomic quest: A personal narrative*, Oxford University Press, 1956.

Corchón, Luis C and Marco Serena, "Contest theory," in "Handbook of Game Theory and Industrial Organization, Volume II," Edward Elgar Publishing, 2018, pp. 125–146.

Ely, Jeffrey C and Balazs Szentes, "Natural Selection of Artificial Intelligence," Technical Report 2023. Working Paper, Northwestern University.

Hirshleifer, Jack, "From weakest-link to best-shot: The voluntary provision of public goods," *Public Choice*, 1983, *41* (3), 371–386.

Konrad, Kai A, *Strategy and Dynamics in Contests*, Oxford University Press, 03 2009.

Rhodes, Richard, *The Making of the Atomic Bomb*, Simon and Schuster, 1986.

Speer, Albert, *Inside the Third Reich*, Macmillan, 1970.

Ulam, Stanislaw M, *Adventures of a Mathematician*, Scribners, 1976.

Yudkowsky, Eliezer, "Artificial intelligence as a positive and negative factor in global risk," in "Global catastrophic risks," Oxford University Press, 2008, p. 308–345.