The IAEA Solution: Knowledge Sharing to Prevent Dangerous Technology Races

Eoghan Stafford* and Robert F. Trager†

July 2022

Abstract

The world appears to be entering an era of heightened great power technological competition in areas such as artificial intelligence. This is concerning because deploying new technologies often involves private benefits and broadly distributed risks. We analyze a dynamic model of a technology competition with negative externalities. Competitors can sometimes reduce risks through a self-enforcing pact in which a laggard agrees not to pursue one kind of technology and in exchange the leader shares other technical discoveries. When the rewards from technological preeminence are high, rivals can only avoid a race if the gap between them is neither too large nor too small. Tech-sharing bargains preserve the laggard's threat to resume racing if the leader reneges, and therefore can work when the leader cannot credibly promise to simply pay the laggard not to race. We further show that tech-sharing bargains do not require perfect monitoring capability by either party.

Preliminary. Please do not cite or distribute without permission.

[*] (corresponding author) Research Scholar, Centre for the Governance of AI; eoghan.stafford@governance.ai

[†] Associate Professor, UCLA Department of Political Science; rtrager@ucla.edu

In the early 1960s, US President Kennedy warned the public to expect a vast increase in the number of nuclear weapons states within a decade.[1] Many believed this expansion would dramatically raise the risk of a nuclear conflict. Yet, the increase did not occur. One reason appears to be the signing of the Non-Proliferation Treaty (NPT), which consists, in part, of a fundamental bargain: nuclear weapons states would share civilian nuclear technology with non-nuclear weapons states in return for commitments not to pursue nuclear weapons. In this article, we analyze whether actors in future technology races might make a similar bargain — some technology sharing in return for some restraint — in order to limit joint risks.

Great powers are increasingly participating in a dangerous and destabilizing race to acquire artificial intelligence (AI) technologies that will give them an edge in military and political competition. China has declared that it seeks to lead the world in AI by 2030.[2] Meanwhile, the recently published Final Report of the National Security Commission on Artificial Intelligence declares that the United States "must win the AI competition that is intensifying strategic competition with China."[3] Other actors are pursuing their own intensive development plans. Such competitions are concerning not only because of the ways they might upend global stability, but also because they limit opportunities for states to cooperate in managing the substantial risks from the development of advanced AI applications.

We model such competition between states, in which research can yield powerful advantages to whoever reaches a major milestone first but, along the way, some new technologies may have negative consequences that spill over to all competitors. We characterize conditions under which race leaders choose to share some of their discoveries with laggards in order to deter participation in the race.

[1] Press Conference, 21 March 1963, *Public Papers of the Presidents of the United States: John F. Kennedy, 1963* (Washington, D.C., U.S. Government Printing Office [GPO], 1964), 280; *New York Times*, 23 March 1963.

[2] State Council of China 2017.

[3] Schmidt et al. 2021, 2.

We find that such entry-deterrence-through-sharing equilibria exist even when actors cannot verify that others are complying with an agreement. In fact, such equilibria are available under some conditions when wealth-transfer equilibria, in which a technology leader "buys off" a laggard, are not. The size of the gap between the research leader and a laggard is a critical factor determining whether sharing equilibria exist. Sometimes, under conditions we explain below, the gap must be neither too small nor too large. Thus, partial sharing of technology advances in return for restraint in pursuing the most dangerous technologies may be a viable solution for some of the governance challenges posed by powerful emerging technologies.

The model we analyze is unlike others in the literature. Many models of arms or technology races are static models with finite, known endpoints (Kydd 1997; Armstrong, Bostrom and Shulman 2016; Naudé and Dimitri 2018), but arms competition takes place over undetermined lengths of time that are themselves endogenous to the strategic process. Without a dynamic framework, we cannot ask whether the threat of mutual defection can enforce welfare-improving equilibria, how actor incentives are expected to change over time, and how these changes affect the possibility of near-term cooperation. Some models employ the repeated games framework (Axelrod 1984; Downs and Rocke 1990; Fearon 2018), which implies an infinite horizon, and thus allow at least for these questions to be asked.[4] These models do not allow for changing stocks of arms or knowledge that persist from period to period, however, which are essential to arms race processes.[5] One arms race paper is an exception (Fearon 2011), but it focuses on Markov perfect equilibrium in which the actors cannot achieve cooperation by threatening to punish uncooperative behavior. Such an ap-

[4] Other models in this substantial literature include Powell 1993 and Jackson and Morelli 2008.

[5] Han et al. (2020) employ an evolutionary approach to understanding safety-research investment tradeoffs. That framework does not model the actors' strategic choices, and instead assumes that a natural selection process can weed out unsafe AI developers. This assumption is in contrast to other formulations of the strategic dilemma, including the one we adopt here, in which there is always some risk when deploying new technologies, and failures affect other players.

proach is valuable in that it can inform about factors that influence actors' incentives, but it does not allow us to investigate the potential effects of self-enforcing agreements, which experience indicates were essential in past competitions. Below, we analyze a dynamic model with a changing stock of knowledge for each of two players who are able to respond to each other's actions. One of our objectives is to demonstrate a partly analytical and partly computational approach that can be applied with relative ease to examine the consequences of a wide range of assumptions beyond the premises we will examine here.

The technology race we describe is also different from other race models, such as those designed to represent patent races and other sorts of technological development races.[6] Unlike these other contexts, actors in our model have the opportunity — and sometimes the incentive — to share knowledge in each period. There is a benefit to research advancements when they occur, in each period, as well as negative externalities from some technological discoveries that turn out to have unintended consequences when first deployed. There is also both a benefit to the actor who is the first to reach a threshold and a cost to the actor who does not. We conceive of this benefit and cost as the result of developing a "transformative artificial intelligence" (TAI), but the same analytical structure applies to other races in which there is a strong incentive to be the first to reach some milestone.

The baseline model reveals that the gap in the players' technological knowledge, which has not been incorporated into similar strategic models before, has interesting effects. When the benefits of small, near-term, or "basic" AI advances are low enough, the leader can persuade the laggard not to compete if they are far enough behind. But when the benefits of basic AI advances are higher and the stakes of being first to TAI are also high, a race can only be avoided if the laggard is close enough to the leader to compel them to make a deal. While reliable compliance monitoring is clearly useful for facilitating agreements between rivals, we show through extensions to the model that technology-sharing agreements can be a viable way to prevent a race even when actors cannot perfectly observe whether each side is

[6] See Langinier and Moschini 2002 for a review of this literature.

complying. We show that such technology-sharing agreements can prevent a race when it is impossible for the leader — or anyone else — to make cash transfers to keep the laggard out of the race. By allowing the laggard to maintain their threat to resume racing, technology sharing allows the leader to credibly commit to upholding their end of the deal.

After discussing the main model and these extensions, we present a case study of the Treaty on the Non-Proliferation of Nuclear Weapons, which illustrates how international technology-sharing agreements can work in practice. The case also demonstrates that such trades can be necessary to convince states to give up the pursuit of a powerful technology.

Approaches to Modeling an AI Arms Race

Advances in AI are having an increasingly significant impact on nations' security. New types of AI are likely to affect many areas of inter-state competition, including disinformation, intelligence analysis, military logistics and decision-making, cyber warfare, and lethal autonomous weapons.[7]

Despite the leading role played by the private sector in developing AI[8] and the fact that AI innovations have been "diffusing rapidly" across borders,[9] competition between states for advantages in AI remains a cause for concern. Not only are governments like China spending billions of dollars annually on AI research,[10] they also have policy tools for monopolizing access to some AI discoveries. States like the U.S. also have the ability to nationalize tech-

[7] See: Ayoub and Payne 2016; Allen and Chan 2017; Thomas 2020; Schmidt et al. 2021. AI can also affect national security indirectly: as automation increasingly transforms economic activity, the invention and diffusion of AI innovations will shape economic growth and therefore the relative resources available to states (Scharre 2021).

[8] Arnold, Rahkovsky and Huang 2020; National Science Board 2022

[9] Schmidt et al. 2021

[10] Acharya and Arnold 2019

nologies that are important to their security.[11] Meanwhile, in other states like China, the government and technology sectors are already tightly linked. All states also retain tools that enable them to slow the diffusion of technological discoveries to other countries, such as controls on exports of technologies and data, and restrictions on international academic collaboration.[12]

Among computer scientists and scholars in cognate fields, a growing community perceives substantial risks to be inherent to development in AI.[13] Zwetsloot and Dafoe (2019) distinguish three categories of AI risks: "misuse", "accident", and "structural." "Misuse" risk is related to the fact that many AI technologies seem to be "dual-use": easily adaptable for malicious purposes their developers didn't intend. For example a commercial drone might turn out to be easy for terrorist groups to modify to carry explosives (Danzig 2018; Zwetsloot and Dafoe 2019). "Accident" risk refers to situations in which an AI application turns out not to achieve its intended goal at all or to have significant negative side effects, such as a self-driving car that fails to detect pedestrians (Kumar et al. 2019; Zwetsloot and Dafoe 2019). Finally, AI may create "structural" risk by changing strategic incentives in dangerous ways. For example, AI algorithms that make it easier to destroy another nuclear-armed state's second-strike capability may create incentives to escalate conflicts or launch preemptive strikes (Moore Geist 2016; Zwetsloot and Dafoe 2019).

Harms from some AI systems can spill over onto many other people besides their designers. In the 2010 "Flash Crash," stock trading algorithms drove a selling frenzy that briefly

[11] Baker 2021.

[12] Paarlberg 2004; Fischer et al. 2021

[13] Other fields of research may be associated with similar risks. Although some of these dangers are speculative, and may be unlikely, they cannot be discounted, because their effects are potentially large. Researchers in biology may threaten humanity by engineering a pathogen that escapes into the environment (Beckstead and Ord 2014). Or, in similar fashion, an industrial process may have poorly understood effects because of insufficient research into its safety. The mining of the oceans' hydrothermal vents is an example of an industrial process that may disrupt Earth's ecosystem (Ahnert and Borowski 2000).

destroyed over a trillion dollars of wealth (Allen and Chan 2017). YouTube's video recommendation algorithm, designed to maximize user engagement, appears to have inadvertently directed users to view extremist content (Rudner and Toner 2021).

Any new technology can have consequences its designers didn't intend. Yet trends in AI toward greater speed, complexity, and integration across a range of sectors suggest it may make it particular hard to anticipate these risks and to correct problems before they inflict substantial damage (Danzig 2018; Zwetsloot and Dafoe 2019; Rudner and Toner 2021).

Unlike the large literature on patent races (Langinier and Moschini 2002), more recent studies ask how technology races can be made *safer* by ameliorating the "all-out" nature of the race so that fewer parties participate or do so with greater care for the general welfare. The literature to date does not, however, analyze strategic contexts with the essential features of these risky technology races. The arms race literature tends to represent races in which there are decreasing returns to investment. If two adversaries build a hundred ships, building one more will have less effect on the balance of power than doing so when each side has a single ship in its fleet. The first ship built has a larger marginal impact on a state's ability to achieve its objectives than the last. Many races of the past had this character in part because they involved scaling up using existing technology. Once the U.S. and Soviet Union both possessed nuclear arms, for instance, the nuclear race was a decreasing-returns context, at least in terms of the direct impact of arms on security as opposed to any impact of arms on status, which may sometimes accrue to the actor with the larger arsenal. Once you can largely destroy your adversary, even after it has largely destroyed you, there's no need for the ability to destroy it a second time; from the material strategic standpoint, the race can stop.

A race to acquire nuclear weapons *technology* has a fundamentally different character. In this case, almost building a weapon has little value. The beginning stages of investment in development produce no direct effects on state capacities or relations. Only once a line is crossed and all the needed breakthroughs have been made is there a discontinuous change in actor capabilities.

The race to acquire TAI and other technologies participates in both of these dynamics, but it more nearly resembles the latter. In the near term, research in AI has immediate benefits across economic sectors, from defense to grading homework assignments. These types of developments might be characterized by decreasing marginal returns, but they are not as severely decreasing as the returns to building an additional nuclear warhead because new technology implies new economic rents and other new capabilities. In the long-term, many believe that progress in AI may be characterized by tipping points. As Good (1966) pointed out a half century ago, computing speeds imply that a general artificial intelligence with the ability to improve itself might quickly achieve abilities that are hard even to comprehend but which would be advantageous to an actor that could control it.[14] Thus, the race may be characterized by dramatically increasing returns, but the time and effort required to produce those returns is unknown. This is one of the factors that makes these technology races importantly different from most arms race models. In a practical sense, these dynamics imply that in such technology races, from the point of view of individual actors, "more is better" to a greater degree than in traditional arms races.[15]

Another difference between an AI race and many arms races of the past is in the degree of domination that winning the race may imply. An actor that controls a TAI could have the ability to develop extraordinary new capabilities, the limits of which are impossible to predict. Even near-term AI progress in technologies like continuous monitoring and processing of information, cryptography, and autonomous weapons could radically alter the balance of power, possibly leading to new instances of domination by some political entities. The

[14] There is no guarantee of an "intelligence explosion" to extreme levels because the returns to investment in improvement may be severely decreasing. As Russell (2019, 143) points out, however, there is no convincing argument that "creating any given level of machine intelligence is simply beyond that capacity of human ingenuity."

[15] See Huntington 1958 on the difference between qualitative and quantitative arms control, and Glaser 2000 for a review.

potential for domination and the degree to which "more is better" mean that an AI race may be more competitive than arms races of the past. The presence of significant global risk makes it potentially more dangerous, especially if those risks are not widely appreciated. There is a strong need therefore to understand opportunities to ameliorate the potentially all-out nature of the race.

This article investigates these questions through a model that builds on several strands of literature. In fact, however, no previous study analyzes a race model that combines the following features: (1) Research progress is stochastic such that the timing of developments and any endpoint, if there is one, is unknown. (2) The actors develop stocks of knowledge or technology as time progresses. (3) The actors make strategic choices throughout. (4) There is a risk of accidents or negative externalities that affect all actors. We know of no model in any discipline that combines these four, and to them, we add one other distinct feature: (5) Actors can choose to share a portion of their research results.

In this article, we study the possibility that race leaders could deter laggards from participating in a technology race through sharing a portion of the fruits of their research. This arrangement allows both players to increase their technological knowledge at the same rate in expectation as if they were researching separately, while investing fewer resources because they are not making redundant discoveries. Potentially more importantly, it reduces the competitive dynamics of the technology race, potentially facilitating greater reflection before risky technologies are deployed and reducing the number of deployments of technologies that turn out to be harmful. Because the leader shares technology it has already tested, the laggard is also able to avoid repeating the leader's mistakes.[16]

From the point of the laggard, this tradeoff — giving up pursuit of certain technologies in return for assistance with others — mirrors the bargain at the heart of the nuclear non-proliferation regime. States that sign on to the Non-Proliferation Treaty receive assistance

[16] Although we do not model investments in AI safety in this paper, a technology-sharing arrangement that induces one player not to compete would also ameliorate pressure to cut corners on investing in safety.

with their civilian nuclear energy programs; in return, they agree not to pursue nuclear weapons. The International Atomic Energy Agency (IAEA) plays an important role in sharing technical knowledge with NPT signatories, helping them not only to expand their peaceful nuclear capabilities, but also to make their nuclear activities safer and more secure, reducing risks to other countries. While this approach has been controversial, with some worrying that civilian assistance has encouraged the spread of weapons technology,[17] the NPT has arguably proven remarkably successful. Although observers at the start of the nuclear age predicted the emergence of dozens of nuclear states within a couple decades, and although approximately 40 countries have the capability to develop nuclear weapons, only 9 have done so.[18] Other factors contribute to this outcome, including alliance nuclear umbrellas, but the carrot of civilian nuclear assistance appears to have been an important factor in convincing countries like Egypt and Ukraine to refrain from developing or keeping nuclear weapons.

In the context of AI research, the approach we study has an important benefit over attempts to buy off actors, namely, that the leader has a continuing incentive to comply. Sharing knowledge produces cooperation in cases where ongoing resource transfers cannot because, in the case of resource transfers, the sides realize that eventually the leader will have no incentive to continue with the transfers. Once the leader is far enough ahead, the laggard cannot race effectively if it decided to do so; therefore, there is no need to continue to buy it off. Realizing in the beginning that the transfers will be short-lived, the laggard will not agree to a deal. Lump sum resource transfers from the leader to the laggard are similarly ineffective unless contracts can be enforced by governing powers because once the laggard has been paid, it can simply continue the race. Sharing knowledge avoids these dilemmas by maintaining the credible threat of the laggard reentering the race.

In the baseline model we describe below, players are fully informed of each other's actions

[17] Bluth et al. 2010; Fuhrmann 2012.

[18] Campbell, Einhorn and Reiss 2005.

and all the players understand the meaning of a “unit” of research effort. We do not study the question of how research gains might be quantified in the real world. Although this question is essential for practical applications, we believe it is also secondary to identifying the strategic incentive conditions under which the actors can escape the all-out race equilibrium. A plausible quantification strategy in an agreement between the parties might be simply to require that the research leader provide the laggard with all research product up to a certain date. Alternatively, particular technological milestones might be identified and disclosed according to a schedule. Other sharing strategies may be preferable, however.

Because players are fully informed in the baseline model, laggards do not have to question whether leaders are sharing their most important findings, and leaders do not have to question whether laggards are cheating on any commitments to refrain from research activities. In two extensions, we show that technology-sharing bargains can still be feasible even when the leader cannot perfectly monitor the laggard’s compliance or vice versa. Indeed, there are additional reasons to suppose that the full-information assumption may be less influential than it might at first appear. From the laggard’s point view, it is probably often the case that the leader can only get away with hiding research gains if the leader is itself not profiting from those gains by turning them into commercial products. It is true that the leader could move closer to TAI without the laggard knowing, but if the laggard cooperates, it has already ceded the possibility of achieving TAI to the leader in any case. From the leader’s point of view, the possibility of cheating by the laggard is not that concerning as long as it can only be done on a small scale. In fact, some cheating could even be beneficial to achieving cooperative equilibria. The ability to cheat a little bit would encourage laggards to participate in agreements that they otherwise would not, and yet it might not be so concerning that it prevents the leader from cooperating. Since cheating on a grand scale would require hiring and resource-acquisition decisions that would likely be visible, the assumption that players cannot cheat on a grand scale without detection may be reasonable, and we show below that it is not strictly necessary. Nevertheless, the degree of verifiability of agreements that is required for cooperation is an important topic for future research.

A Model of an AI Race

We model the pursuit of AI by two states as a two-player infinitely repeated game with complete information. Players accumulate basic AI knowledge by investing in research. If a player accumulates enough knowledge, they make a breakthrough discovery we refer to as "transformative AI" (TAI).[19] In implementing new basic AI technologies for the first time, a player sometimes causes accidents or other unintended consequences that inflict costs on both players. Finally, a player with more advanced technology can share some of their knowledge with the less advanced player.

Moves

At each point in the game, players have a stock of basic AI knowledge ($A_{i,t} \in \{0, 1, ..\}$ for each player i in period t). There is a gap G between the players' initial knowledge levels, with Player 1 leading ($A_{1,1} - A_{2,1} = G \geq 1$).[20]

At the beginning of each period players choose either to invest in research ($R_{i,t} = 1$) or not investing ($R_{i,t} = 0$). Players incrementally add to their stocks of basic AI knowledge through research, but discoveries occur probabilistically. If a player researches, there is a probability $\nu \in (0, 1)$ that they make a discovery in that period ($X_{i,t} = 1$) and otherwise their research yields no advances that period ($X_{i,t} = 0$). The probability of a discovery is 0 in a period when a player does not research.[21]

[19] This term has been defined in very different ways by other scholars. (See Gruetzemacher and Whittlestone 2019 for a review.) We use it here in the very general sense of a technological breakthrough that results in a qualitative shift in the strategic context.

[20] Rather than representing all AI technology, this knowledge should be thought of as representing knowledge in a more particular domain within AI, in which each discovery builds directly on the previous one.

[21] The assumption that new technological discoveries happen randomly is important because it implies that

Research can represent direct R&D carried out by a state or R&D by private actors such as firms. In the latter case, the model would apply to actors whose technology one state has a comparative in accessing (e.g. through export controls or informal pressure). In either case, it makes sense to speak of a gap between states in their access to technology.

There is some threshold level of knowledge $\overline{A}$ at which a player discovers TAI. This threshold is greater than the leader's initial knowledge level, but the players do not know what the threshold is until one of them reaches it. Players share prior beliefs about the value of $\overline{A}$, which for tractability, we assume is a geometric distribution: conditional on the most advanced player having reached knowledge $A_{i,t} = a$ without discovering TAI, the probability that the next discovery will reach the TAI threshold ($\overline{A} = a + 1$) is $\theta \in (0, 1)$. We assume that TAI represents a qualitative shift: players have no further moves in any periods after one player reaches the threshold.

In each period that a player researches, there is a probability $\phi \in (0, 1)$ of that player causing a "failure." As we discuss in the next section, a failure imposes costs on *both* players, not just the one who caused it. Both players can cause failures in the same period and we assume that failures occur independently. Also, a player can make a beneficial discovery and cause a failure in the same turn.

Failure events reflect the idea of "failure modes" discussed by machine learning researchers.[22] When prototypes are first deployed, they sometimes turn out to cause unanticipated harms to the global economy, international security, the climate, etc. As previously noted, these harms may come in the form of "accidents," vulnerability to "misuse," or "structural" effects on the global strategic environment (Zwetsloot and Dafoe 2019). The harms may be due to design flaws, but could also be caused by the manner in which states use a technology or they way they regulate (or fail to regulate) its use by private actors.

there is a chance for a player who is behind to eventually overtake the player who is currently in the lead even if both players are researching.

[22] See, for example Kumar et al. 2019.

At the end of each period, after the players observe whether either made a discovery and whether either caused a failure, if one player has accumulated more basic AI knowledge than the other, that player chooses an amount of knowledge to share with the less advanced player (T_t). The leader can choose not to share any knowledge ($T_t = 0$), to share all of their knowledge ($T_t = (A_{i,t} + X_{i,t}) - (A_{j,t} + X_{j,t})$) or any amount in between. A player in the lead or tied after research discoveries are revealed in t starts the next period with knowledge $A_{i,t+1} = A_{i,t} + X_{i,t}$, while a lagging player has $A_{j,t+1} = A_{j,t} + X_{j,t} + T_t$.

Technology-sharing in the model captures a broad range of ways that states can transfer technology, whether providing training to scientists in other countries or simply abstaining from technology export controls. The knowledge can be about not only technical designs, but also ways of operating the technology safely, or policy approaches to manage its social risks.

A player bears no failure risk from any knowledge that the other player transfers to them. This assumption captures the idea sharing technical designs, implementation strategies, and regulatory approaches the leader has already successfully implemented allows the laggard to avoid repeating any mistakes the leader made in the process of discovery.

To summarize the sequence of moves in each period:

- Each player chooses whether to research ($R_{i,t} \in \{0, 1\}$).
- Players learn whether each has made a discovery ($X_{i,t} \in \{0, 1\}$) and whether each has caused a failure ($F_{i,t} \in \{0, 1\}$).
- Players learn whether either player has reached TAI ($\overline{A} = \max\{A_{i,t} + X_{i,t}\}$) or not ($\overline{A} > \max\{A_{i,t} + X_{i,t}\}$).
- If one player has more knowledge than the other ($A_{i,t} + X_{i,t} > A_{j,t} + X_{j,t}$), the more advanced player chooses an amount of knowledge to share ($T_t \in \{0, ..., (A_{i,t} + X_{i,t}) - (A_{j,t} + X_{j,t})\}$).

Payoffs

Each player gets a per-period payoff based on their AI knowledge at the beginning of the period. We follow models in the economics literature in assuming that short-term benefits increase when a player has more AI knowledge than others do (Budd, Harris and Vickers 1993). Players split the benefit of the knowledge they have in common, while a leading player gets all the benefit of their extra knowledge that the lagging player lacks. Each unit of knowledge produces a total benefit $\alpha > 0$. Suppose that $A_{i,t} \geq A_{j,t}$. Both players get $\alpha/2$ from each of the first $A_{j,t}$ units of knowledge. The leader gets an additional α for each of the $A_{i,t} - A_{j,t}$ units of knowledge that only they have discovered.[23]

The player who reaches TAI ($A_{i,t} = \overline{A}$) first gets a benefit $\tau > 0$ in each subsequent period, while the losing player pays a cost $\kappa > 0$ each period. (If both players reach the TAI threshold in the same period, we assume that each has a 50% chance of winning.) Players therefore have two motives for maintaining a technological lead for two reasons: to get more benefit from their basic AI knowledge and to get to TAI first. These conflicting interests represent the idea that having access to innovations in AI allows a state to compete more effectively with its rivals.

Each player suffers a loss $\lambda > 0$ for each failure caused by *either* player.

A player who invests in research ($R_{i,t} = 1$) pays a cost normalized to 1. This cost can be thought of as representing expenditures on R&D (by the state or companies under its influence) as well as the expected costs of failures that do *not* impose externalities on other states.

Players have a common discount factor $\delta \in (0, 1)$.

[23] Players continue to get these payoffs from their cumulative knowledge in each period forever, even after the race ends.

Equilibria of the Baseline Model: Racing and Technology-Sharing

If players do not condition on each other's past actions, they simply race to develop more advanced AI. We distinguish two different kinds of races that can occur in equilibrium. In a Perpetual Race equilibrium, both players invest in research until one of them discovers TAI ($R_{1,t} = R_{2,t} = 1 \ \forall \ t$ such that $A_{1,t}, A_{2,t} < \overline{A}$). In a Limited Race equilibrium, a player will drop out (permanently) if they fall too far behind. That is, there is some gap $g^{\bigstar} \geq 1$ such that, if $A_{i,t} - A_{j,t} \geq g^{\bigstar}$, then $R_{j,t} = 0$.[24]

However, if players condition on what has happened in previous periods, there can be a cooperative equilibrium in which they avoid a race. In such an equilibrium, the leader shares basic AI technology with the laggard, who in turn refrains from pursuing TAI. Each period, if the leader discovers a new unit of knowledge ($X_{1,t} = 1$), they share one unit of less advanced knowledge with the laggard ($T_t = 1$), maintaining a constant gap G between the two players' knowledge levels.[25] The only exception is the final period, when the leader reaches TAI ($X_{1,t} = 1$ and $A_{1,t} + 1 = \overline{A}$), at which point they do not share their last basic AI discovery ($T_t = 0$).

This deal is enforced by a grim trigger. If the leader doesn't share any technology after making a discovery ($X_{1,t} = 1$ and $T_t = 0$), or if the laggard does any research ($R_{2,t} = 1$), then the game enters a race: players play either the Limited or Perpetual Race subgame equilibrium in every subsequent period.[26]

[24] There can also be a "Never Race" equilibrium, in which players don't research regardless of how far behind or ahead they are.

[25] That is, at the beginning of each period, the leader and laggard have respective knowledge levels $A_{1,t}$ and $A_{2,t}$, with $A_{1,t} = A_{2,t} + G$. If the leader makes a discovery, their knowledge level in the next period will be $A_{1,t+1} = A_{1,t} + 1$. Since they share a unit of knowledge with the laggard, the latter will have $A_{2,t+1} = A_{2,t} + 1$.

[26] For all parameter values, at least one of the three non-cooperative equilibria exists – Perpetual Race, Limited Race, or the equilibrium in which players never invest. Therefore, in presenting our results from the baseline

The leader benefits from this cooperative arrangement because they are guaranteed to reach TAI first, while the laggard benefits by gaining access to a growing stock of knowledge without having to invest in research themselves. In addition, both players benefit from a reduction in the expected number of failures, since only the leader is doing research, while the laggard is only implementing designs that the leader has verified work as intended.[27]

Mind the Gap: Results of the Baseline Model

Intuition might suggest that a dangerous race can be avoided if one of the competitors is far enough ahead. Let's call this idea the Hopeless Laggard effect: If the lagging player is far enough behind, and so their chance of winning the race is small enough, they would accept a tech-sharing bargain. If they are even further behind, it isn't worth it for them to race even in the absence of a tech-sharing bargain. This Hopeless Laggard effect sometimes happens, but not always. In our model, we find that, in some cases, a race can only be avoided if the gap between the players is neither too small nor too *big*. We call this the Goldilocks effect. In these cases, the benefits of basic-AI advances are so large that the laggard is always willing to invest in research, no matter how far behind they are. The leader is only willing make a technology-sharing deal if the laggard is close enough to pose a credible threat of winning a race. The laggard would make a deal, rather than try to overtake the leader, as long as they are not *too* close.

model, we assume that players always follow one of these non-cooperative equilibria if they do not follow the tech-sharing equilibrium.

[27] The essential features of this kind of technology-sharing are that the lagging state gives up its pursuit of some powerful technology in exchange for assistance from the leader with some other technology, and that the transfer of technical knowledge from one state to another reduces the shared risks associated with the technology. In practice, such an agreement might allow a lagging state to do research in related areas, so long as that research did not bring them too close to the prohibited transformative technology. Indeed, in the case of the IAEA itself, non-nuclear weapons states can research civilian nuclear technologies, so long as they do not try to develop nuclear weapons.

In this section, we present conditions for when a large or a small knowledge gap is needed to prevent a race. When the benefits from basic-AI advances are small, a race can be avoided if the gap between the players is large enough (the Hopeless Laggard effect). When the benefits from basic AI are in a middle range, a race can only be avoided — if at all — when the gap is in a middle range (the Goldilocks effect). We present sufficient conditions for the Goldilocks effect. Among these conditions is that the stakes of acquiring TAI must be sufficiently high. Finally, if the benefits of basic AI are too high relative to the other parameters, such as the players' subjective evaluations of the risks of the race, it becomes impossible to avoid a race, no matter how small or large the gap is.

These results follow from how the benefits of basic AI determine what the players do in the absence of an agreement. A Perpetual Race only occurs if the benefits of basic AI are so large that it is worth investing in research even as a player's probability of winning the race to TAI approaches zero. For any values of the other parameters, there is always some such threshold $\overline{\alpha}$ such that, if the basic-AI benefit is above $\overline{\alpha}$, Perpetual Race is an equilibrium, while below it, Limited Race is an equilibrium.

If the basic-AI payoff is low enough that Limited Race is an equilibrium, the Hopeless Laggard effect *must* hold. If the laggard starts off far enough behind, the race will be over before it's begun: even though the leader will not share technology, the laggard will not invest in research.

The Hopeless Laggard effect *can* hold even if the basic-AI payoff exceeds the threshold $\overline{\alpha}$, so that a Perpetual Race would occur in the absence of a cooperative agreement. The larger the knowledge gap, the lower the chance that the laggard will ever catch up, which reduces the leader's incentive to make a deal. However, the leader has another reason to make a deal besides guaranteeing their victory in the TAI race: avoiding the spillover costs of failures the laggard would cause if they researched. If these costs are significant, the leader can credibly commit to sharing technology despite an arbitrarily large gap. The laggard may not be willing to make a deal if the gap is small enough to give them a significant shot at catching up, but for a large enough gap, they will be willing to cooperate.

By contrast, if the basic-AI payoff is too high relative to the expected cost of failures, then for some large enough gap the leader would defect to not sharing discoveries. Let us call the threshold basic-AI payoff above which the leader prefers to defect for a large enough gap $\overline{\overline{\alpha}}$. If the expected cost of failures is high enough, then $\overline{\alpha} < \overline{\overline{\alpha}}$: there is a range of basic-AI payoffs in which a Perpetual Race would occur in the absence of a technology-sharing deal, but for a large enough gap, both players will agree to a deal.[28]

If and only if the basic-AI payoff is below either $\overline{\alpha}$ or $\overline{\overline{\alpha}}$, then the Hopeless Laggard effect holds. If the basic-AI payoff is higher than both these thresholds, then the Hopeless Laggard effect does not hold: for all sufficiently large gaps, the equilibrium outcome is a Perpetual Race.[29]

Figure 1 illustrates the different equilibrium outcomes that occur when the knowledge gap becomes very large.[30] The Hopeless Laggard effect holds only in the regions where no race occurs if the gap is sufficiently large. In the "No Race, No Sharing" region, the basic-AI payoff is low enough that the laggard doesn't race if they are far enough behind, even though the leader does not share technology. For basic-AI payoffs above the "No Race, No Sharing" region, no matter how large the gap in players' knowledge levels, the laggard prefers to race if the leader does not share technology. In the "No Race, Sharing" region, the basic-AI payoff is nonetheless low enough relative to the high expected cost of failures that the leader shares technology to keep the laggard from conducting research. In the "Race" region, the basic-AI payoff is so high that the leader can't commit to sharing technology, so both players carry

[28] The pace of discoveries (ν) must also be high enough for this range to exist.

[29] For low expected failure costs, the first threshold is binding: a large gap can only prevent a race if the benefits of basic AI alone do not justify investment for a laggard who is too far behind. For higher expected failure costs, the second threshold is binding: a large gap can prevent a race if either a distant laggard wouldn't race *or* a distant laggard would race but the leader would share technology.

[30] This figure holds fixed $\delta = 0.95$ and $\nu = 0.25$.

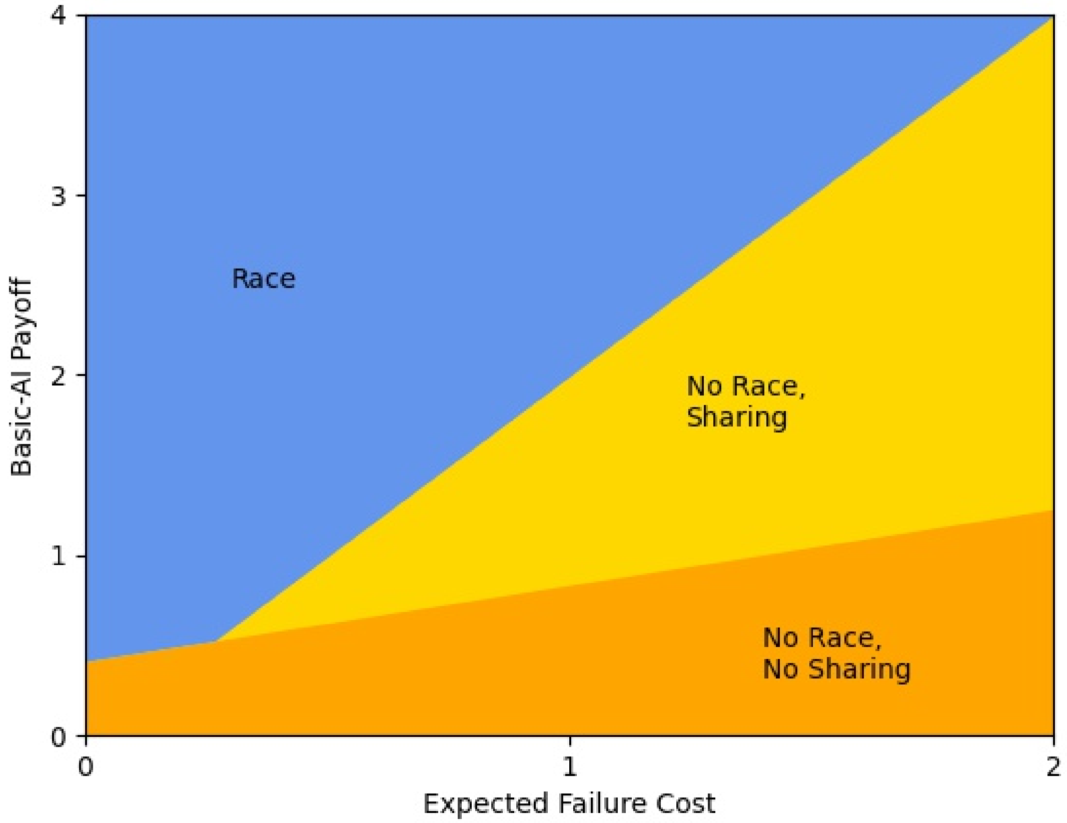

Figure 1: Equilibrium outcome for large knowledge gaps

out research, resulting in a Perpetual Race.[31]

Proposition 1 provides the necessary and sufficient condition for the Hopeless Laggard effect. The proposition highlights the importance of the returns from basic AI. When these returns are low, the players can avoid dangerous races when one is sufficiently ahead. The method of avoiding the race could be either through a sharing agreement or through the

[31] The threshold $\overline{\alpha}$ is the upper bound of the "No Race, No Sharing" region, and it is upward sloping because the higher the expected cost of failures, the more valuable advances must be for a distant laggard to invest in research perpetually. The threshold $\overline{\overline{\alpha}}$ is the upper bound of the "No Race, Sharing" region. It slopes upward because the higher the expected cost of failures, the greater the cost to the leader of triggering a race by defecting.

laggard unilaterally dropping out.[32]

Proposition 1 *(**Hopeless Laggard Effect**):*

There is a threshold $\alpha^\star > 0$ such that, if and only if the benefit of each basic-AI advance (α) is at most $\alpha^\star$, then there is a threshold $\overline{G} \geq 1$ such that, if the initial gap (G) is at least $\overline{G}$, there exists an equilibrium in which the lagging player never researches ($R_{2,t} = 0$ for all periods t).

If, on the other hand, the basic-AI payoff is so large that the Hopeless Laggard effect doesn't hold, it may nonetheless be possible to avoid a race if the gap is *not* too large. In a close race, the leader may be willing to make a technology-sharing bargain because the chance of the laggard overtaking them is very high. Depending on the other parameters, there may be a middle range in which the gap is small enough that the leader would cooperate and large enough that the laggard will cooperate too. This is the Goldilocks effect mentioned above.[33]

Figure 2 illustrates a case where the Goldilocks effect is in operation.[34] Each line represents the difference between a player's expected utility from cooperating and their expected utility if they defect, triggering a race. If the difference is positive, the player would cooperate; if it's negative, they prefer to defect. The leader's relative payoff from cooperating declines as the gap between the players increases: as the leader becomes more likely to get to TAI first even if a race occurs, cooperation becomes less appealing. The reverse is true for the laggard, whose relative payoff from cooperating increases in the gap size. When the

[32] We present and prove a full statement of Proposition 1 in Online Appendix A.1.

[33] There are cases where cooperation is possible even if the laggard is only one unit of knowledge behind. In particular, such cases occur when the stakes of TAI are in a middle range. However, the implication that any small enough gap is sufficient for cooperation in such cases may be an artifact of using discrete knowledge levels. If we were to model knowledge levels as continuous, we might find that there is always some range of gaps close to zero at which the laggard would be unwilling to cooperate.

[34] This figure holds fixed $\delta = 0.95$, $\nu = 0.4$, $\theta = 0.1$, $\lambda\phi = 0.25$, $\alpha = 0.9$, $\tau = 1$, and $\kappa = 4$.

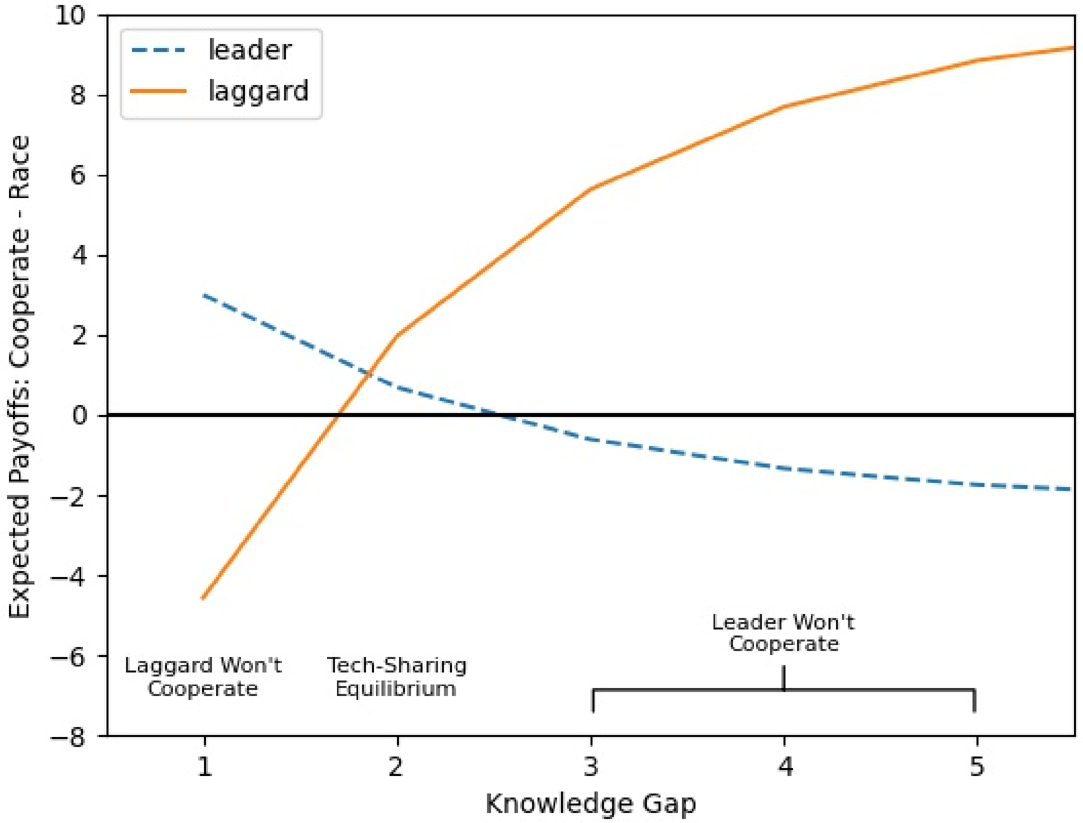

Figure 2: Incentive to cooperate by player (high-stakes TAI)

knowledge gap is 2, both players would cooperate. If the gap is only 1, the laggard would defect. If the gap is 3 or more, the leader would defect.[35]

A second proposition summarizes a set of sufficient conditions for the Goldilocks effect:[36]

Proposition 2 *(**Goldilocks Effect**): If the following conditions are true:*

- *the probability of making a basic-AI discovery each period (ν) is high enough*

[35] For some other parameter values, the middle range in which cooperation is possible includes more than one gap size.

[36] In Online Appendix A.2, we present a full statement of Proposition 2 and explain how we derived it on the basis of the results established in Proposition 1 and numeric computations. We describe our numerical methods in Online Appendix B.

- *the basic-AI payoff (α) is in a middle range, the lower bound of which is at least $\alpha^{\star}$*

- *the probability that any given discovery leads to TAI (θ) is low enough*

- *the stakes of the race for TAI ($\tau + \kappa$) are high enough*

then there is some range of gap sizes defined by $1 \leq G' \leq G''$ such that:

- *if the initial gap $G \in \{G', ..., G''\}$, the tech-sharing equilibrium exists*

- *and otherwise, the tech-sharing equilibrium does not exist and a Perpetual Race occurs ($R_{1,t} = R_{2,t} = 1$ in all periods t until one player discovers TAI).*

There is an upper bound on the basic-AI benefit because both players' incentives to defect are increasing in that parameter. The higher the basic-AI benefit, the smaller the range of knowledge gaps in which both players would cooperate. If it is too high, there is no gap at which both players will cooperate.

The last condition, about the benefit of winning and cost of losing the TAI race, is perhaps the most surprising. Why would raising the stakes of TAI improve the feasibility of cooperation? A key point for understanding this result is that there is always some gap at which the laggard would cooperate. As the gap gets bigger and their chance of catching up approaches zero, they would be strictly better off in expectation receiving technology from the leader, rather than bearing the costs (from investment and additional failures) of making advances for themselves. Given the lower bound on the basic-AI payoff, the leader will only cooperate if the gap is small enough. The bigger the difference between the payoffs of winning and losing the race to TAI, the higher the maximum gap at which the leader would cooperate. If the TAI stakes are large enough, the range of gap sizes in which the leader would cooperate overlaps the range in which the laggard would cooperate.

Figure 3 is analogous to Figure 2, but with a much smaller difference in payoffs between

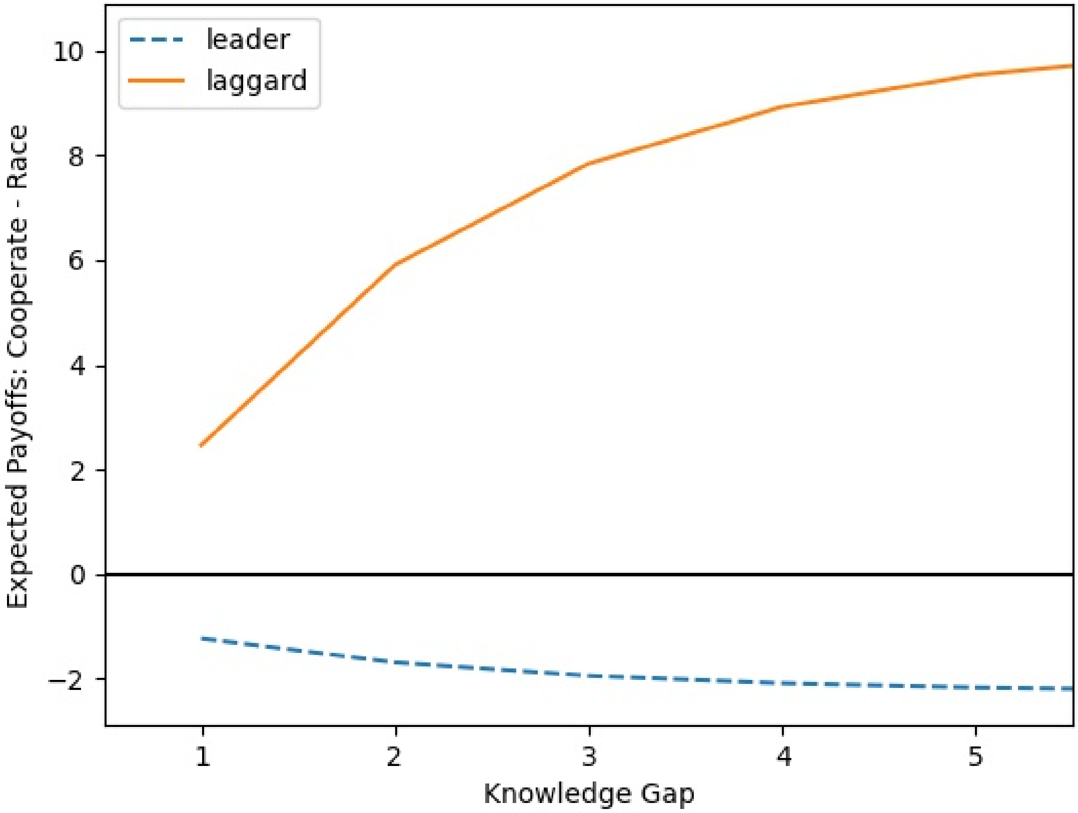

Figure 3: The players cannot cooperate through sharing because TAI is low stakes

winning and losing the race for TAI.[37] With the stakes of TAI lowered, the laggard has less to gain from racing and would cede TAI to the leader no matter how small the gap is. However, a race has become less costly for the leader, and so no matter how small the gap, they prefer to defect and not share any discoveries they make.

In short, when the stakes of TAI are high enough and the benefits of basic AI are in a middle range, the players can only avoid a race if the knowledge gap is small enough.

[37] This figure holds the parameters fixed at the same values as Figure 2 except that the cost of losing the race to TAI is lower: $\kappa = 1$.

Extensions

Incomplete Information

A simplifying assumption of our model is that actors have complete information. In the real world, it may be difficult for rival companies, states, or other competitors to monitor each other's research activities or discoveries. The laggard might cheat on the agreement by secretly carrying out research or the leader might cheat by concealing their technological advances rather than sharing them with the laggard. If either actor had significant doubts about their ability to monitor the other, they might be unable to agree to an otherwise mutually beneficial technology-sharing agreement.

Such difficulties with monitoring compliance would certainly make a technology-sharing bargain more difficult in practice. However, we analyzed two extensions of the model with incomplete information and found that technology sharing can be feasible even in the absence of perfect monitoring.

In the Hidden Research extension, the laggard has the option of carrying out research in secret, so the leader cannot monitor the laggard's compliance with the technology-sharing deal. If their research goes well, they can potentially beat the leader to TAI, and if it does not, they continues to receive technology from the leader, ensuring that they never fall further behind than they were when the agreement began.

We assume that making extensive use of any knowledge they have acquired that the leader has not yet shared with them would be observable, so the laggard can choose temporarily to not deploy those discoveries. Thus, while their research is hidden, they do not get any benefit from their knowledge beyond the level the leader has shared. Once they choose to carry out their research openly, they receive the full benefits of their knowledge. There is still a risk, however, that the laggard will cause a failure when they test new discoveries, and if a failure happens, the leader infers that the laggard was carrying out research, triggering the end of the technology-sharing agreement.[38]

Through numerical computation, we find that there are indeed some parameter values for which the laggard can profitably cheat in this way. Anticipating that the laggard will cheat, the leader will not agree to share technology in the first place. So, there are some cases where technology sharing is possible in the baseline model but not in this extension.

However, we find that there are other parameter values for which the laggard cannot gain from secretly researching, so a technology-sharing agreement can be possible even when the leader is unable to verify that the laggard is not conducting research. For instance, if in expectation it will take many periods to reach TAI, or if failures are very likely or costly, the possibility of winning the TAI race is outweighed by the cost of investment and increased failure risk.[39]

We also analyze a Hidden Discoveries extension, in which the laggard cannot observe whether the leader is complying with the deal. Specifically, the laggard does not observe whether the leader makes a discovery each period. The exact technology-sharing arrangement we analyze in the baseline model is clearly not feasible in this version of the model, since the laggard cannot condition on whether the leader shares an old technology every time they make a new discovery. Instead, we check the existence of an equilibrium in which the leader shares one unit of technological knowledge with the laggard every K periods. If the leader

[38] We assume that the probability and cost of a failure is the same when the laggard is secretly testing new designs as when players openly implement them for the first time. It might be more realistic to assume that the risks are smaller since the laggard would presumably have to conduct testing on a smaller scale if they are to keep it secret. However, relaxing our assumptions in this way shouldn't qualitatively effect the result, because the laggard would still have to tradeoff the chance of overtaking the leader with the cost of investing in research and at least some increased failure risk. Indeed, even if there was no chance of causing a failure while conducting secret research and therefore no chance of getting caught, for some parameter values the laggard would still be better off giving up their shot at TAI in order to avoid paying for their own research, since they will receive the leader's basic AI discoveries either way.

[39] See Online Appendix C.1 for a formal description of the model and our numerical methods for analyzing it.

ever fails to transfer a unit of technology in a scheduled period or if the laggard ever carries out research, the bargain ends and the players follow whichever non-cooperative subgame equilibrium exists.

Since discoveries are stochastic and the leader may make more than one discovery in any given K periods or make none, the gap may grow or widen over time. As a result, there is always a chance that the deal will collapse. The gap can close completely as a result of the leader failing to make discoveries in multiple periods. If there is no gap in a period when the erstwhile leader is scheduled to transfer technology to the other player, they will not have any technology to share that the laggard doesn't already have. Because the laggard can't observe whether the leader has been making discoveries, if the leader does not share technology, the laggard can't tell whether the leader is reneging or has simply been unlucky.

Nonetheless, we find that this kind of technology-sharing agreement is an equilibrium for some parameter values. Numerical analysis suggests that a tech-sharing equilibrium exists as long as the probability and benefits of each basic-AI discovery are in a middle range and the probability of any given discovery leading to TAI is sufficiently low. Thus, both of these extensions provide a strong test for the feasibility of tech-sharing agreements: in each one, monitoring is not just imperfect, but non-existent for one of the players.[40]

Wealth Transfers

So far, we have only considered one way that competitors could avoid a race: by sharing technology. In practice, there are other agreements they might strike. One possibility is that the leader might simply *pay* the laggard not to research, without transferring technology.[41] Such a solution would be more robust against unanticipated events that might motivate the laggard to rejoin the race. As the technological gap between leader and laggard grows, the

[40] See Online Appendix C.2 for a formal description of the model and our numerical methods for analyzing it.

[41] Another way the leader might compensate the laggard would be to allow them to use the technology without sharing the knowledge of how the technology works, e.g. through an API.

risk of a race breaking out would decline.

We created an extended version of the model in which the leader can transfer wealth, rather than technology, to the laggard in exchange for the laggard refraining from research. We find that a wealth-sharing agreement is indeed self-enforcing for some parameter values. There are cases in which both a technology-sharing agreement and a wealth-sharing agreement are feasible, but there are also cases in which only a wealth-sharing agreement is feasible.

However, we also find cases in which a technology-sharing agreement is possible, but a wealth-sharing agreement is not. Because the technology gap grows under a wealth-sharing agreement, the prospect of a race becomes less costly to the leader and so the maximum payment they would make steadily declines. If a Limited Race is a subgame equilibrium, the leader could eventually stop making payments altogether without triggering a race, and so the laggard would have an incentive to defect and begin racing before the gap reached that point.[42] Anticipating the laggard's defection, the leader would defect when the gap was even smaller, and so on, so that players would not agree to the deal in the first place in equilibrium.[43]

These cases, in which tech-sharing agreements are feasible but wealth-sharing agreements are not, occur when advances in basic AI occur rapidly but the discovery of TAI is a long way off in expectation. Wealth-sharing deals unravel if the probability of a discovery each period is too high, because the laggard anticipates with a high degree of confidence when the gap is about to widen to the point where the leader would stop making payouts. Because a rapid

[42] However, even when the non-cooperative equilibrium is a Limited Race, a wealth-sharing agreement can still be feasible if discoveries occur with a low probability, so that even when the leader is just one unit away from stopping transfers, the payments continues for such a long time in expectation that the laggard wouldn't want to defect.

[43] There are also cases in which a wealth-sharing equilibrium exists but the leader eventually stops paying even though the resulting race is perpetual. In these cases, while a technology-sharing agreement could guarantee that the laggard would never race, a wealth-sharing agreement could not.

pace of basic AI discoveries does not affect the gap between the players under a technology-sharing agreement, the laggard would cooperate so long as the probability that any one discovery leads to TAI is low, because the agreement lasts a long time in expectation.[44]

The NPT: A Case Study

The fact that there are only nine nuclear weapons states today — rather than the dozens Kennedy warned of — is likely in part due to the global nuclear nonproliferation regime that is centered around the Treaty on the Non-Proliferation of Nuclear Weapons. Also known as the Nuclear Non-Proliferation Treaty (NPT), this international agreement was a technology-sharing bargain in which nuclear-armed states provided civilian nuclear technology to other states, and the latter agreed in turn not to acquire nuclear weapons themselves.[45] The case confirms that dynamics similar to those described in the model can be important elements in the governance of dangerous new technologies. It also demonstrates that such trades can be necessary to convince states to give up the pursuit of a powerful technology. Technology sharing was acceptable to the US and Soviets because some states were close to acquiring nuclear technology. It was acceptable to those other states because they got things in return, including civilian nuclear technology and security guarantees. The deal thus reconciled the goals of (1) enabling less technologically-advanced powers to reap the benefits of civilian nuclear energy technology, (2) reducing the global risks of nuclear energy programs, and (3) discouraging beneficiary states from pursuing nuclear weapons.

Some scholars reject the idea that the NPT was meant to be a bargain to trade civilian nuclear tech for giving up nuclear weapons. Swango (2014) argues that when the United States ratified the treaty in 1968, it was not committed to providing assistance to states

[44] See Online Appendix D for a formal description of the model and our numerical methods for analyzing it.

[45] Müller 2005; Cirincione 2007; Findlay 2012 The NPT requires NWS themselves to give up nuclear weapons eventually, albeit without specifying a timeframe. However, possible limitations on states' use of an advanced tech (represented by TAI in our model) is beyond the scope of our model.

that renounced nuclear weapons. At the same time, the US was willing to share nuclear technology with some states that remained outside the NPT. We show, however, that the promise of civilian nuclear assistance and cooperation was essential to the selling of the agreement to the powers who would be denied the possibility of pursuing nuclear weapons.

Civilian and military applications of nuclear science play a similar role to basic and transformative AI, respectively, in our model. This is so for several reasons. First, nuclear technology is dual-use: the technical knowledge required in nuclear energy and other civilian applications can also make it easier for a state to develop nuclear weapons.[46] Second, even the civilian applications of nuclear technology pose risks that spill across borders, from accidents to the theft of nuclear material by terrorist groups,[47] risks that are heightened if states do not have access to lessons learned from others' experience.[48] Third, since much of even civilian nuclear technology is not publicly available, we can meaningful speak of gaps between different states in terms of their access to advanced nuclear technology.[49]

There are also some differences between the nuclear case and the strategic context we describe in the model. A significant one is that, unlike TAI in our model, nuclear weapons technology already existed at the time that the NPT was created. This is essentially a simplification of the modeling context in which the leader achieves the major advance (TAI) immediately; the fundamental strategic dynamics are similar. Another difference between the case and model is in the number of actors. While our model is a two-player game, the NPT is a global multilateral agreement. However, it codifies a bargain between two groups: the original five nuclear weapons states (NWS) and the world's many more non-nuclear weapons states (NNWS). Larger numbers of states add complexity, but we believe that some fundamental strategic problems are evident in both contexts.

[46] Müller 2005; Pilat and Busch 2015

[47] Pilat and Busch 2015; Findlay 2011

[48] Findlay 2011; Taebi and Mayer 2017

[49] Kroenig 2009

The NPT emerged from negotiations between the United States and the Soviet Union aimed at limiting the spread of nuclear weapons to other countries.[50] The two superpowers reached agreement by the end of 1966 that the treaty should prohibit the transfer of nuclear weapons technology to any states that did not already have it.[51] However, many of the NNWS objected to the two-tiered system this proposed treaty enshrined: a permanent division between the nuclear weapon "haves" (the US, the USSR, the United Kingdom, France, and China) and "have-nots" (the rest of the world).[52] Swango (2014, 211) — who, as noted earlier, rejects the claim that the NPT can be seen as a technology-sharing bargain — remarks: "Non-aligned states and the superpowers' own allies argued they could not join a discriminatory system that permitted the superpowers to possess nuclear weapons unless they received something in exchange for adhering."

Among the countries objecting to the US-Soviet draft treaty were several nations that a 1958 US National Intelligence Estimate judged to be capable of developing nuclear weapons within a decade.[53] Getting such countries to join the NPT was a "particular concern for the superpowers," but these close laggards "realized that signing the treaty could have economic and security drawbacks" and insisted that the treaty guarantee their access to "nuclear technology for peaceful purposes" (Hilborne 2012, 254).

In response to the demands of the NNWS, Washington and Moscow revised the NPT, starting in the second half of 1967. The UN General Assembly approved the final version of the treaty in June 1968 only after two further articles had been added guaranteeing, as Bourantonis (1997, 356) summarizes, "the sharing of benefits from the peaceful application of nuclear energy" with states that agreed not to pursue nuclear weapons. Articles III

[50] Cirincione 2007, 30

[51] Goldschmidt 1980, 75

[52] Müller 2005, 46 Glenn Seaborg, a former chair of the US Atomic Energy Commission later observed: "The non-nuclear countries were not about to accept without resistance a pact that they believed to be highly discriminatory against them" (quoted in Cirincione 2007, 31).

[53] Goldschmidt 1980, 74; Cirincione 2007, 27

and IV not only permitted NNWS to engage in civilian nuclear activities but explicitly ensured the transfer of knowledge related to civilian nuclear technology. While the NNWS had to submit to verification by the IAEA that they were not developing nuclear bombs, Article III required that the verification measures must "avoid hampering the economic or technological development of the Parties or international co-operation in the field of peaceful nuclear activities." Article IV creates a positive obligation of signatories to share peaceful nuclear technology: "All the Parties to the Treaty undertake to facilitate, and have the right to participate in, the fullest possible exchange of equipment, materials and scientific and technological information for the peaceful uses of nuclear energy."[54] Whereas the original US-Soviet negotiations focused exclusively on banning the proliferation of nuclear weapons, "It was not until the rest of the world got involved with the negotiations that its most contentious provisions — those calling for sharing nuclear technology for peaceful purposes and for the gradual elimination of the nuclear weapon states' arsenals — were debated and finally accepted."[55]

Assessing the overall impact of the NPT on nuclear proliferation is complex. Some evidence suggests that states which receive technical assistance with civilian nuclear activities are more likely to pursue and ultimately acquire nuclear weapons.[56] However, one of these studies also shows that signing up for the NPT reduces the chances states will start a nuclear weapons program.[57] And Gibbons (2020, 282) finds that it is precisely the prospect of receiving civilian nuclear technology that has persuaded some states to join the NPT, because: "After the risks of nuclear assistance became well-known following India's nuclear explosion in 1974, most major suppliers conditioned their assistance on recipients joining nonproliferation agreements." Whether or not the US intended, at the time of the NPT's

[54] The text of the NPT is available at: www.un.org/disarmament/wmd/nuclear/npt/text

[55] Cirincione 2007, 30

[56] Kroenig 2009; Fuhrmann 2009; Brown and Kaplow 2014

[57] Fuhrmann 2009

ratification, to use civilian nuclear assistance to induce individual states to join the treaty, the offer of such assistance seems to have been crucial to both the passage of the the treaty and the large number of states that subsequently joined it.

Thus, like the players in the technology-sharing equilibrium of our model, the signatories of the NPT know that NNWS *could* use their knowledge of civilian nuclear technology to pursue a weapons program.[58] But they calculate that most NNWS will have a strong incentive not to jeopardize their continued access to peaceful nuclear technology — and that many of them *would* pursue weapons in the absence of such assistance and cooperation.

Information sharing with the NNWS serves a second purpose beyond inducing states to renounce nuclear weapons: it improves the safety and security of nuclear facilities, reducing risks that spill across borders. In the last three decades the IAEA has played a growing role in providing that assistance, advising states on preventing accidents at nuclear reactors and securing nuclear fuel against theft.[59] In particular, the IAEA's Global Nuclear Safety and Security Network supports "the transfer of knowledge from countries with mature nuclear energy programmes to countries that have only just started to embark on such programmes."[60] As in our model, knowledge-sharing under the nuclear non-proliferation regime reduces the global risks of nuclear energy production in less developed countries.

The NPT case illustrates how the results of our model can occur in practice. A bargain to share one kind of technology can prevent competition for another, related technology. The negotiations that led to the NPT also illustrate the Goldilocks Effect. The nuclear-armed powers were willing to concede civilian nuclear assistance because several states posed a

[58] A 2004 UN report notes: "Almost 60 states currently operate or are constructing nuclear power or research reactors, and at least 40 possess the industrial and scientific infrastructure which would enable them, if they chose, to build nuclear weapons at relatively short notice *if the legal and normative constraints of the Treaty regime no longer apply*" (emphasis added; quoted in Cirincione 2007, 106).

[59] Findlay 2011; Taebi and Mayer 2017

[60] International Atomic Energy Agency, "Global Nuclear Safety and Security Network (GNSSN)", accessed May 20, 2022, https://www.iaea.org/services/networks/global-nuclear-safety-and-security-network

credible threat of developing their own atomic weapons in the near future. Most of the non-nuclear-armed countries were willing to join the NPT because they were not *too* close and assistance with nuclear energy was a valuable inducement.

Discussion

For most of the Cold War, the United States and the Soviet Union continued to deploy increasing numbers of nuclear warheads atop intercontinental ballistic missiles. In the late 1980s, this nearly continuous increase was replaced by a nearly continuous decline on both sides. This change coincided with an agreement between the sides, the Strategic Arms Reduction Treaty (START I), which was negotiated in the 1980s and signed by George H. W. Bush and Mikhail Gorbachev in 1991. How far did the sides reduce their arsenals of deployed weapons? Down to about 6,000 on each side, the number allowed by the treaty. Agreements have an important role to play in arms control, but the strategic context determines which can be self-enforcing in the absence of a common governing authority. The results of the models presented here confirm that even in the highly adversarial context of technology competition, and even without the mitigating factor of the decreasing returns from weapons technology investment in the nuclear race, agreements between adversaries that reduce the level of arms or technology competition are possible.

These results contrast with those that derive from static race models. The role of information about each other's behavior, for instance, is different. Dynamic models allow agents the possibility of punishing adversaries who fail to adhere to agreements. Implementing such strategies requires a certain amount of information, and this implies that information can play a welfare-increasing role that is absent from static models. While perfect information is not required for cooperative equilibria, the absence of any expectations about an adversary's behavior precludes agreements that require conditioning on it.

Another conceptual issue, however, is whether the actors know exactly what is required to achieve TAI. This sort of information is subtly different from knowledge of relative position in the search for TAI. The welfare decreasing effect of information in the Armstrong, Bostrom and Shulman 2016 model, for instance, comes not from information about relative position directly, but from the heightened competition that erupts when actors know exactly what they need to do to outdo each other, and from the assumption that doing slightly more leads to all the benefits with certainty. At least in the near term, however, since it remains unknown what is required to produce TAI, this seems not to be a faithful representation of the strategic context. This is another reason to expect information about actors' actions and knowledge to have a more beneficial effect, at least in the short term when actors do not know the precise set of steps required for TAI.

Such static models, and dynamic models that are analyzed through Markov perfect equilibria, place the emphasis on the material strategic context. The dynamic game that we analyze puts the emphasis instead on the social world that actors could create. It allows us to ask what agreements would be self-enforcing in the absence of a common authority to compel compliance.

We stated at the outset that this paper would provide a baseline model, and indeed, there are many extensions that would be useful for understanding these dynamics. We might analyze larger numbers of players or the dynamics of a model with the possibility of conflict or sabotage. One could also study the degree of resource investment in order to understand, for instance, how investment changes as players are closer and farther from each other in the AI knowledge race (Bimpikis, Ehsani and Mostagir 2019), and how this impacts safety dynamics. One could study learning dynamics as players observe aspects of each other's progress, such as how players might be given an incentive to invest or not based on other players' success. Innovations may provide investment and racing incentive because they represent proof of the possible. This modeling direction would enable the generation of hypotheses about when to expect secrecy versus product development that employs recent innovations.

A promising direction for future work is to model a tradeoff between advancing quickly and advancing safely. This research direction includes models of NGO, governmental, and intergovernmental actors who seek to increase race safety. Such actors may have the paradoxical problem of not being able to credibly threaten to take actions that risk safety. They may not wish, for instance, to threaten to race unsafely themselves unless other actors race safely. On the other hand, in a situation with multiple potential racers, an altruistic actor could offer to help the safest actor with the goal of influencing the behavior of all. In the best-case equilibrium, the threat from the altruistic actor would deter other actors from racing unsafely without the altruistic actor even having to commit any resources at all. The threat of committing resources if another actor deviated could be enough. This would mirror some models of lobbying in which lobbyists achieve their objectives without committing any funds — because of the threat to fund political adversaries. How these dynamics would interact with different relative positions in the research race is something for modeling to investigate.

Although considerations about the social and global risks of advanced technologies to societies at large motivate this study, our model only allows direct analysis of the welfare of the potential participants in a technology race. The balance of the benefits and risks to societies is likely to vary considerably across different applications of AI, not to mention other advanced technologies, and has crucial implications for whether a technology-sharing agreement that is mutually beneficial to its participants would be in the interests of affected third parties. For the most dangerous technologies, such agreements may not go far enough: a total ban on research in some areas may be preferable for humanity as a whole. For other technologies, the benefits to third parties may be so large that an agreement which reduces competition and delays advances in expectation may be detrimental to global welfare.

Conclusion

Some interesting policy considerations derive from the results of the model presented above. For instance, partial sharing of technology may be a tool to prevent more dangerous forms of proliferation. Intergovernmental organizations or other third-party actors could facilitate technology-sharing bargains between states by creating an institution that serves the function of the International Atomic Energy Agency. Such an institution would share knowledge with certain states in return for commitments not to engage in proscribed research, development, implementation and use. They could also help to compel behavior through sanctions regimes similar to those in place in the nuclear field to prevent the spread of technologies. This is useful not just because of the direct effect of limiting the spread of dangerous technologies, but also because it preserves the value of sharing those technologies in return for actions that increase public welfare.[61]

The analysis here does not lead to firm policy conclusions, however. In spite of our findings that perfect information is not required for sharing equilibria, some capacity for compliance verification — or at least detection of cheating on a grand scale — will often be beneficial, even necessary, in the real world. Some actors will be willing to take on more risk that adversaries are cheating. As in the nuclear non-proliferation regime, minor powers under a major power's nuclear umbrella will take risks in forgoing technological development that major powers themselves would not. The right policy approaches for different technologies may vary, and they on the balance of benefits and risks to global welfare, not only to the particular governments competing most fiercely over it.

Contemporary political discourse and popular media often frame AI research as an inevitably cutthroat race for global dominance. That rhetoric runs the risk of becoming a

[61] While the scope for imposing rules on major powers is limited, international institutions can reduce the transaction costs of agreements between states, making it worthwhile for them to enter into technology-sharing agreements that would not otherwise be feasible (Keohane and Martin 1995).

self-fulfilling prophecy.[62] Our findings provide reason for optimism that nations can find ways to reduce conflict in their pursuit of advanced AI. At the same time, those findings demonstrate why we cannot take such cooperation for granted.

[62] Scharre 2019; Sherman 2019; Kerry et al. 2021

References

Acharya, Ashwin and Zachary Arnold. 2019. Chinese Public AI R&D Spending: Provisional Findings. Technical report Center for Security and Emerging Technology.
URL: *https://cset.georgetown.edu/publication/chinese-public-ai-rd-spending-provisional-findings/*

Ahnert, Ahmed and Christian Borowski. 2000. "Environmental risk assessment of anthropogenic activity in the deep-sea." *Journal of Aquatic Ecosystem Stress and Recovery* 7(4):299–315.

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

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

Arnold, Zachary, Ilya Rahkovsky and Tina Huang. 2020. Tracking AI Investment: Initial Findings From the Private Markets. Technical report Center for Security and Emerging Technology.
URL: *https://cset.georgetown.edu/publication/tracking-ai-investment/*

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

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

Baker, James E. 2021. "A DPA for the 21st Century: Securing America's AI National

Security Innovation Base." *CSET Policy Brief* .
URL: *https://cset.georgetown.edu/publication/a-dpa-for-the-21st-century*

Beckstead, Nick and Toby Ord. 2014. Managing Existential Risk from Emerging Technologies. In *Annual Report of the Government Chief Scientific Advisor.* pp. 115–120.

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

Bluth, Christoph, Matthew Kroenig, Rensselaer Lee, William C. Sailor and Matthew Fuhrmann. 2010. "Civilian Nuclear Cooperation and the Proliferation of Nuclear Weapons." *International Security* 35(1):184–200.

Bourantonis, Dimitris. 1997. "The Negotiation of the Non-Proliferation Treaty, 1965–1968." *The International History Review* 19(2):347–357.

Brown, Robert L. and Jeffrey M. Kaplow. 2014. "Talking Peace, Making Weapons: IAEA Technical Cooperation and Nuclear Proliferation Journal of Conflict Resolution." *Journal of Conflict Resolution* 58(3):402–428.

Budd, Christopher, Christopher Harris and John Vickers. 1993. "A model of the evolution of duopoly: Does the asymmetry between firms tend to increase or decrease?" *The Review of Economic Studies* 60(3):543–573.

Campbell, Kurt M., Robert J. Einhorn and Mitchell B. Reiss, eds. 2005. *The Nuclear Tipping Point: Why States Reconsider Their Nuclear Choices.* Washington, D.C.: Brookings Institution Press.

Cirincione, Joseph. 2007. *Bomb Scare: The History & Future of Nuclear Weapons.* New York: Columbia University Press.

Danzig, Richard. 2018. "Technology Roulette: Managing Loss of Control as Many Militaries

Pursue Technological Superiority.”.
URL: *https://www.cnas.org/publications/reports/technology-roulette*

Downs, George W and David M Rocke. 1990. *Tacit Bargaining, Arms Races, and Arms Control*. Ann Arbor, MI: University of Michigan Press.

Fearon, James D. 2011. “Arming and Arms Races.” *Annual Meetings of the American Political Science Association, Washington, DC* .
URL: *https://www.researchgate.net/publication/254063608*

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

Findlay, Trevor. 2011. *Nuclear Energy and Global Governance: Ensuring Safety, Security and Non-proliferation*. New York, NY: Routledge.

Findlay, Trevor. 2012. Unleashing the Nuclear Watchdog: Strengthening and Reform of the IAEA. Technical report The Centre for International Governance Innovation.
URL: *https://www.cigionline.org/publications/unleashing-nuclear-watchdog-strengthening-and-reform-iaea*

Fischer, Sophie-Charlotte, Jade Leung, Markus Anderljung, Cullen O’Keefe, Saif M. Khan, Stefan Torges, Ben Garfinkel and Allan Dafoe. 2021. AI Policy Levers: A Review of the U.S. Government’s Tools to Shape AI Research, Development, and Deployment. Technical report Centre for the Governance of AI, Future of Humanity Institute, University of Oxford.
URL: *https://www.governance.ai/research-paper/ai-policy-levers-a-review-of-the-u-s-governments-tools-to-shape-ai-research-development-and-deployment*

Fuhrmann, Matthew. 2009. “Spreading Temptation: Proliferation and Peaceful Nuclear Cooperation Agreements.” *International Security* 34(1):7–41.

Fuhrmann, Matthew. 2012. *Atomic Assistance: How "Atoms for Peace" Programs Cause Nuclear Insecurity*. Ithaca, NY: Cornell University Press.

Gibbons, Rebecca Davis. 2020. "Supply to Deny: The Benefits of Nuclear Assistance for Nuclear Nonproliferation." *Journal of Global Security Studies* 5(2):282–298.

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

Goldschmidt, Bertrand. 1980. "The Negotiation of the Non-Proliferation Treaty (NPT)." *IAEA Bulletin* 22(3/4):73–80.

Good, Irving John. 1966. Speculations concerning the first ultraintelligent machine. In *Advances in Computers*, ed. Franz L. Alt and Morris Rubinoff. Vol. 6 New York, NY: Academic Press pp. 31–88.

Gruetzemacher, Ross and Jess Whittlestone. 2019. "Defining and Unpacking Transformative AI." *unpublished manuscript* .
URL: *https://www.researchgate.net/profile/Jess-Whittlestone/publication/337702892*

Han, The Anh, Luis Moniz Pereira, Francisco C. Santos and Tom Lenaerts. 2020. "To regulate or not: a social dynamics analysis of the race for AI supremacy." *arXiv preprint* .
URL: *https://arxiv.org/pdf/1907.12393.pdf*

Hilborne, Mark P. 2012. The Non-proliferation Treaty: Foundation of Disarmament Policy. In *Handbook of Nuclear Proliferation*, ed. Harsh V. Pant. New York, NY: Routledge pp. 251–260.

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

Jackson, Matthew O and Massimo Morelli. 2008. "Strategic militarization, deterrence and wars." *Available at SSRN* .
URL: *https://papers.ssrn.com/sol3/papers.cfm?abstract*$_i$*d* = 1081775

Keohane, Robert O. and Lisa L. Martin. 1995. "The Promise of Institutionalist Theory." *International Security* 20(1):39–51.

Kerry, Cameron F., Joshua P. Meltzer, Andrea Renda, Alex Engler and Rosanna Fanni. 2021. Strengthening international cooperation on AI: Progress report. Technical report Brookings.

URL: *https://www.brookings.edu/research/strengthening-international-cooperation-on-ai/*

Kroenig, Matthew. 2009. "Importing the Bomb: Sensitive Nuclear Assistance and Nuclear Proliferation." *Journal of Conflict Resolution* 53(2):161–180.

Kumar, Ram Shankar Siva, David O. Brien, Kendra Albert and Snover Jeffrey Viljöen, Salomé. 2019. Failure Modes in Machine Learning Systems. Technical report arXiv.

URL: *https://arxiv.org/abs/1911.11034*

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

Langinier, Corinne and GianCarlo Moschini. 2002. The Economics of Patents: An Overview. CARD Working Papers 02-WP 293.

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

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

Müller, Harald. 2005. Peaceful Uses Of Nuclear Energy And The Stability Of The Non-Proliferation Regime. In *Effective Non-proliferation: The European Union and the 2005 NPT Review Conference*, ed. Darryl Howlett, John Simpson, Harald Müller and Bruno Tertrais. European Union Institute for Security Studies pp. 43–62.

National Science Board. 2022. Science and Engineering Indicators 2022: The State of U.S. Science and Engineering. Technical report National Science Foundation.
URL: *https://ncses.nsf.gov/pubs/nsb20221*

Naudé, Wim and Nicola Dimitri. 2018. "The race for an artificial general intelligence: Implications for public policy." *AI & Society* 35(2):367–379.

Paarlberg, Robert L. 2004. "Knowledge as Power: Science, Military Dominance, and U.S. Security." *International Security* 29(1):122–151.

Pilat, Joseph F. and Nathan E. Busch. 2015. Introduction. Nuclear proliferation: A future unlike the past? In *Routledge handbook of nuclear proliferation and policy*, ed. Joseph F. Pilat and Nathan E. Busch. New York, NY: Routledge pp. 1–11.

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

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

Russell, Stuart. 2019. *Human Compatible: Artificial Intelligence and the Problem of Control*. Penguin Books.

Scharre, Paul. 2019. "Killer Apps: The Real Dangers of an AI Arms Race." *Foreign Affairs* 98(3):135–144.

Scharre, Paul. 2021. "Artificial Intelligence: Debunking the AI Arms Race Theory." *Texas National Security Review* 4(3):121–132.
URL: *https://tnsr.org/2021/06/debunking-the-ai-arms-race-theory/*

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

URL: *https://www.nscai.gov/2021-final-report*

Sherman, Justin. 2019. "Stop calling artificial intelligence research an "arms race".".

URL: *https://www.washingtonpost.com/outlook/2019/03/06/stop-calling-artificial-intelligence-research-an-arms-race/*

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

Swango, Dane. 2014. "The United States and the Role of Nuclear Co-operation and Assistance in the Design of the Non-Proliferation Treaty." *The International History Review* 36(2):210–229.

Taebi, Behnam and Maximilian Mayer. 2017. "By accident or by design? Pushing global governance of nuclear safety." *Progress in Nuclear Energy* 99:19–25.

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

Zwetsloot, Remco and Allan Dafoe. 2019. "Thinking About Risks from AI: Accidents, Misuse and Structure.".

URL: *https://www.lawfareblog.com/thinking-about-risks-ai-accidents-misuse-and-structure*