On Decisive Strategic Advantage

L. C. R. Patell

Abstract

The possibility of developing advanced AI capabilities that confer a decisive strategic advantage may lure the United States and China into an accelerating race for technological dominance. That race could usher in a new epoch of strategic competition. While some authors have begun to model the dynamics of these races, there is a wealth of strategic literature from the cold war that has yet to be applied to US AI national security strategy. This paper sketches the co-evolution of history and theory throughout the cold war, with attention to Schelling's dictum that uncertainty can be a strategic tool. We proceed to argue that existing models of US AI national security strategy fail to endogenize uncertainty in this way. In response, we construct a model in which states can signal their capabilities and probe their opponents' by engaging in minor conflicts. While the model is rudimentary, it demonstrates that making uncertainty strategically manipulable can yield unexpected conclusions and thereby help to illuminate the strategic landscape.

Introduction

> The most spectacular event of the past half century is one that did not occur. We have enjoyed sixty years without nuclear weapons exploded in anger".
>
> – Thomas Schelling, 2005

During the Cold War, the combined might of American cognition crafted theory and policy that contributed to this spectacular event. It is time to turn that concerned effort towards advanced AI and the specter of global conflict that looms in its wake. In this decade or the next, we may bear witness to the creation of AI systems that unleash new, devastating military technologies upon the world. These technologies may transform the nature of warfighting, and their very possibility may lure the United States and China into an accelerating race for technological dominance.

Work has begun: some authors have started to model the dynamics of an AI arms race. These models, however, barely scratch the surface of what can be wrung from the strategic literature on the Cold War. Many of the deficits in these accounts can be remedied by closer attention to history and theory.

This paper aims to push forward the application of Cold War-era strategic thinking to AI national security strategy. In doing so, it remains an early foray into a rich theoretical landscape. We aim not to provide the final word, but instead to refine existing models and to draw out a host of questions that merit further investigation.

1 Historical Background

The history of the Cold War is marked by the coevolution of military reality and strategic theory.

The atomic bombings of Hiroshima and Nagasaki transformed warfare. Bernard Brodie grasped the nature of this transformation: "from now on the chief purpose of the military establishment must be to avert wars rather than to win them".[1]

Yet deterrence – convincing an adversary that aggression is self-defeating – did not immediately become the organizing *telos* of military strategy. Possessing nuclear weapons does not by itself guarantee stability. Indeed, once the Soviets conducted their first test of the bomb in 1949 and the nuclear arms race was on, deterrence began to look tenuous. "The Delicate Balance of Terror" (Wohlstetter, 1958) and "The Nature and Feasibility of War and Deterrence" (Kahn, 1960) demonstrate the fragility of deterrence: it depends on the reciprocal belief that each side can absorb a first strike and deliver devastating retaliation. Small shifts in force posture, warning time, and command-and-control could upset that balance and invite preemptive attacks. Both Wohlstetter and Kahn go into excruciating detail: they canvas the possibility that the threat of tens of millions of Russians dead may be insufficient to deter a first strike.[2]

Deterrence, therefore, is a dynamic equilibrium – something that must be engineered, inspected, and repaired – not necessarily a natural resting point.

Thomas Schelling helped supply the analytic scaffolding necessary to validate these claims. "Prospectus for a Reorientation of Game Theory" (Schelling, 1958) recasts political conflict as sequential bargaining games. As a result, commitment devices – anything that credibly locks an actor into a future response – can be just as important as raw firepower. Schelling's reorientation emphasizes that shifting the information environment can be a strategic lever in its own right.[3]

The Strategy of Conflict broadens this picture.[4] Schelling introduces three ideas that still sit at the foundation of deterrence theory. First, focal points: when payoff matrices under-determine equilibrium choices, actors gravitate

[1] Brodie and Dunn (1946).

[2] Kahn asks: "how many American dead would we accept as the cost [of retaliating against a Soviet first strike]". Kahn finds that his fellow citizens place the cost between 10 and 60 million Americans dead (1960, p. 15).

[3] A perhaps obvious, but profound point – one that Schelling stresses in his foreword – is that he analyzes non-zero-sum games, and stresses that any conflict that is not to the death is not strictly zero sum.

[4] Schelling (1960).

towards certain salient solutions. Second, brinkmanship: the deliberate creation of small probabilities of catastrophe – "the threat that leaves something to chance" – as a bargaining chip. Third, tacit bargaining: coordination can emerge without explicit communication, provided each side can read the other side's constraints and incentives. Signals can be just as important for strategic stability as treaty signatures.[5]

Schelling's work is rich: this sketch merely scratches the surface. Nevertheless, we can draw out two essential modeling imperatives.

1. Model uncertainty.
2. Model the manipulation of risk – not all accidents are exogenous shocks.

Working in a game-theoretic setting induces a prior question: how should we model what states want? More precisely, we need to make some modeling choices about the payoffs that states get from different outcomes. One natural answer here stems from the realist tradition in international relations – particularly the canonical work of Morgenthau (2001), Waltz (1979), and Mearsheimer (2001) – which models states as egoistic actors in an anarchic system. Realism can provide a structural fulcrum for modeling a bargaining situation; these models can, at least, serve as a first approximation of the conflict under investigation. Strictly speaking, however, a game-theoretic approach is broader than realism: it is consistent with modeling states with arbitrary payoffs. Realism may downplay domestic politics, regime types, and ideological motives, all of which could shape the bargaining payoffs in more detailed models.

As demonstrated by the early analysis of deterrence, the regime we now call "mutual assured destruction" did not arise the moment the skies of Hiroshima ignited. It emerged in three intertwined phases: technical adequacy, doctrinal codification, and legal entrenchment.

[5] The analysis of signaling and information design has flourishing in the International Relations and Economics literatures. See, for example, Fearon (1958), Fearon (1995), Fearon (1997), and Kamenica and Gentzkow (2011). The notion that this territory has been relatively underexplored in the context of AI and conflict informs Patell (forthcoming a) and Patell and Guest (forthcoming a).

Deterrence, as Wohlsetter and Kahn establish, requires achieving technical adequacy: survivable, devastating second strike capabilities, as well as the capacity to detect a first strike. The Polaris ballistic-missile submarine, the Minuteman solid-fuel ICBM in hardened silos, and dispersal schemes for Strategic Air Command bombers all but ensured that at least one American retaliatory mechanism would survive a first strike. Parallel investments in communications technology increased the likelihood that retaliation orders would be received even after attempts to decapitate the U.S. military. These technical advances, combined with their Soviet counterparts, preserved the mutual vulnerability necessary for MAD to be viable.

Robert McNamara's Defense Department articulated a capability required to "deter potential aggressors" called "assured destruction; i.e., the capability to destroy the aggressor as a viable society, even after a well-planned and executed surprise attack on our forces".[6] Mutual assured destruction emphasizes the realization that symmetrical vulnerability stabilizes the international stage, rather than sheer superiority. Politically durable doctrine crystallized only after technical advances delivered unambiguous constraints: the survivable triad. Yet it is implausible to wait until technologies advance to begin crafting national security strategy; for that reason, attempting to model the situation – including the veil of uncertainty – can provide direction.

Beginning with the Partial Test Ban Treaty in 1963, the nuclear powers began to legally entrench stability. They attempted to ossify the group of states with access to nuclear weapons through the Non-Proliferation Treaty, forced superpowers to limit nationwide missile defense through the Anti-Ballistic Missile Treaty, and instituted a series of technical obligations through SALT I and II. International agreements helped stabilize a regime that technology and theory had jointly pushed us towards: no state could rationally pursue a first-strike strategy.

2 Decisive Strategic Advantage

AI may constitute a significant shock to strategic stability. Here, we consider the possibility that states will develop superintelligence (ASI): an AI system

[6] U.S. Department of Defense (1966); McNamara (1967).

that is superior to all human beings in all cognitive domains.[7] Such a system may have capabilities that expand the frontier of warfighting in both the cyber and kinetic domains. Many commentators claim that a state that unilaterally develops ASI would wield a decisive strategic advantage (DSA): the ability to confidently defeat the rest of the world combined.[8] According to these commentators, ASI could yield a DSA by neutralizing adversaries' nuclear arsenals. Yet, despite their growing prevalence, these claims lack rigorous support. Aside from intuitions that ASI would be unprecedentedly powerful, there has been little analysis of the specific technical pathways through which such a system would disrupt nuclear stability.

Coates (forthcoming) and Winter-Levy and Lalwani (2025) take an important step by interrogating the claim that advanced AI could undermine nuclear deterrence. Coates breaks down the claims that would need to be true for a state to obtain an AI-enabled DSA and how plausible those pathways are. He identifies three primary pathways:

1. **Defeat.** Possessing the offensive capability to completely destroy or disable an adversary's second strike capabilities.

2. **Defend.** Possessing the defensive capability to otherwise neutralize the intended effects of an adversary's second strike capability.

3. **Dominate.** Other pathways – such as economic influence, psychological warfare, or sabotage – that yield a DSA without directly interacting with nuclear weapons.

The core upshot of Coates' argument is that an AI race is marked by significant uncertainty. Uncertainty as to whether states will be able to obtain a DSA at all; uncertainty about whether your current level of AI capabilities

[7] Some commentators reject the possibility of superintelligence outright. For the purposes of argument, we assume that ASI is possible, and attempt to work through the strategic implications that result. We note that it seems premature to confidently assert that ASI is not possible; the precautionary principle suggests that we should begin crafting appropriate strategies and mitigations even if the advent of ASI is uncertain. As long as ASI remains in the possibility space, as we demonstrate in the next section, you can model skepticism by modulating the payoffs in a game-theoretic setting.

[8] The concept of decisive strategic advantage conferred by unilateral possession of ASI stems from Bostrom (2014) and has become central in the discourse on AI and international security.

can provide a DSA; uncertainty about whether a supposed DSA will actually scupper your adversary's strike capabilities; and uncertainty about the capabilities of your opponent.[9]

The weight of this uncertainty may seem daunting. We can, however, begin to model it as part of a signaling game. Before constructing a model, however, we will examine precedent models of U.S. AI national security strategy. In doing so, we will demonstrate that these models fail to capture the uncertainty latent in the strategic situation adequately.

3 AI and National Security Strategy

There have been several attempts to develop U.S. national security strategy on the basis of the possibility of a DSA.

Aschenbrenner (2024) implicitly models the strategic situation as a prisoner's dilemma and advocates for the US to race towards ASI. He writes that pursuing an international treaty is "fanciful", because "breakout" from "any arms control equilibrium" is "too easy: the incentive (and the fear that others will act on this incentive) to race ahead with an intelligence explosion, to reach superintelligence and decisive advantage, too great".[10] Aschenbrenner justifies this claim by evoking the distinction between unstable and stable arms control. While 1980s arms control targeted a stable equilibrium, in which arsenals thousands large sustained MAD, whereas "zero nukes wouldn't be a stable equilibrium". ASI arms control would ostensibly be similarly unstable: "if mere months of lead on AGI would give an utterly decisive advantage [...] a rogue upstart or treaty breaker could gain a huge edge by secretly starting a crash program; the temptation would be too great for any sort of arrangement to be stable".[11] Aschenbrenner assumes that ASI will have god-like powers that enable a DSA, but neglects several sources of uncertainty. As we demonstrate in the next section, that uncertainty may make a significant difference to strategy.

[9] States will also have higher-order uncertainties – uncertainties about uncertainties – about their adversary's perspective.

[10] Aschenbrenner (2024), p. 137.

[11] ibid.

Katzke and Futerman (2024) leverages some explicit game-theoretic modeling to push back against Aschenbrenner's strategic picture. They claim that the strategic situation is characterized by a trust dilemma, such that states should pursue greater trust and cooperation. This conclusion follows from a crucial modeling choice: they claim that states are defensive realist actors, and that they do not gain from breakout. Moreover, they claim that a trio of risks – loss of control, great power conflict, and internal instability – should lead mutual cooperation (a treaty) to be significantly preferred to racing by both the US and China.

Aschenbrenner's argument amounts to the claim that states obtain immense gains from unilateral dominance. Crucially, this assumption is strictly consistent with the assumption that states can be described with defensive realist utility functions: they may see their survival as maximally secured by world domination. But if breakout is incentivized, then the trust dilemma threatens to unravel. A cooperative equilibrium, however, can be sustained in a model that is ambiguous between different payoffs when states know that they are playing a repeated game.

Katzke and Futerman unequivocally assume that developing ASI before your opponent deals them a significant loss. While, in principle, those calculations may include uncertainty, the uncertainty enters the model exogenously – and, hence, these games fall afoul of Schelling's imperatives.

Hendrycks et al. (2025) attempt to invoke the stability-inducing dynamics of MAD via "Mutual Assured AI Malfunction", or "MAIM". They claim that the threat of succumbing to an ASI-enabled DSA will lead states to sabotage their adversary's projects, and that this dynamic will mutually deter states from racing towards ASI. As Rehman et al. (2025), Abecassis (2025), Wildeford and Delaney (2025), and Patell (forthcoming b) argue, this analogy fails. Without suppressed modeling assumptions, the dynamics of MAIM simply do not mirror MAD's exercise in comparative risks. Nevertheless, these dynamics make progress towards models that adequately capture the uncertainties and comparative risks at stake in a race towards ASI.

While not game-theoretic, Amodei (2024) portrays his vision of a desirable future – one that involves an "eternal 1991" in which the U.S. and its allies are "able to parlay their AI superiority into a durable advantage". Latent in this durability is the notion of obtaining – and then *using* – a DSA to defang or pacify America's adversaries. This sort of vision may be one that prompts America's adversaries to view its pursuit of ASI as an existential threat to their security and ambitions for the future; this perception of threat is exactly the cause of preemptive conflict in Hendrycks et al. (2025).

Barnett and Scher (2025) – introducing the Machine Intelligence Research Institute (MIRI's) view of the strategic landscape – remedy a problem from "The Manhattan Trap". While Katzke and Futerman equate "racing" and building a "Manhattan project", Barnett and Scher correctly identify that building a US national AI project does not rule out the possibility of pursuing international cooperation. Instead, "pivoting away from a National Project" is possible.[12]

Barnett and Scher (2025) also write that "a decisive strategic advantage would allow the US to quash any competing AI projects and potentially force other nations to submit to a US-led international regime". This claim is true by definition; the salient question is whether or not the US *would* obtain a DSA through racing. While they note that the DSA strategy "depends on several unstable conditions", they do not explore the strategic ramifications of this uncertainty; instead, they simply reject it because it "is extremely dangerous".[13]

The possibility of using a national project to build leverage that the US can bring to the negotiating table constitutes a central crux in the strategic debate. Katzke (unpublished) argues that a stronger US lead would increase distrust between the U.S. and China, undermining the possibility of cooperation. This argument mirrors Boudreaux et al. (2025), who suggest that the failure of the Baruch plan largely stemmed from a lack of trust between the negotiating parties. At the same time, however, game-theoretic analysis suggests that China's best alternative to a negotiated agreement – its "BATNA" – should be worse the greater the US's lead. If Katzke is right, then the pivot strategy that wrings an international agreement from a US

[12] Barnett and Scher (2025). Patell and Guest (forthcoming b) elaborate on the non-identity between building a National Project and racing, before going on to argue that MIRI misses the possibility of building a National Project and engaging in cooperation in parallel.

[13] Barnett and Scher (2025).

national AI project may be ill-advised. This disagreement – including the crucial role played by distrust – is one that might be fruitfully investigated in a game-theoretic setting that endogenizes uncertainty; we postpone that analysis to future work.

Armstrong et al. (2016) attempts to model race dynamics between differing AI development teams. They argue that "danger is minimized when each team is ignorant of the AI-building capabilities of every team" – that is, when there is maximal uncertainty about your adversary. While this uncertainty does not concern DSAs, it still constitutes a surprising result. Emery-Xu et al. (2024) build on this model by incorporating greater uncertainty and again find that the "no information" equilibrium is safer than one involving public or private information. These conclusions are sometimes taken to demonstrate that states should not divulge any information – or signal at all – in the race setting. Yet that inference is spurious: the models do not endogenize signaling or information sharing, and so they do not yield definite conclusions about signaling and strategic transparency.[14]

4 A Signaling Model

We model two rival states who are competing to develop powerful AI capabilities. Each has some endowment of resources w that leads them to obtain a level of capabilities (Date 1). This process is stochastic, insofar as states do not know *ex ante* whether or not their development trajectories will merely perpetuate the status quo (which we calibrate to a capability level $K_i = 0$) or lead to superintelligence.

Before any major war, each state can initiate a minor conflict: they publicly demonstrate a level of AI capabilities in the theatre of war that constitutes a credible lower bound (Date 2). Intuitively, this conflict is more costly the more powerful and evenly matched the two sides' capabilities are.

Finally, each state decides whether or not to launch a major attack. A major war promises an immense prize – control over the future – to whoever holds a DSA. Peace leads both sides to split control over the future evenly.[15] If

[14] Patell (forthcoming a) attempts to remedy this gap in an information design setting.

[15] In more complex models, peacetime control can be split in more nuanced ways.

neither side has a DSA, however, and they go to war, catastrophe results.

4.1 Primitives

Two states $i \in \{1, 2\}$.

Date 1 (private capabilities).

Each state i has budget $w_i > 0$. To begin, let

$$w_i = w_j$$

Nature draws capabilities

$$K_i \;=\; w_i + \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}(0, \sigma^2) \text{ i.i.d.}$$

Each state privately observes K_i.

Decisive strategic advantage (DSA). State i has a decisive strategic advantage if, and only if:

$$K_i \;\geq\; K_j + T,$$

where $T > 0$ is the decisive threshold; neither state has a decisive strategic advantage otherwise.

Date 2 (minor conflict).

Each state i publicly chooses an attack $y_i \in [0, K_i]$, where $y_i = 0$ corresponds to no attack. In this way, attacking may create a *credible lower bound* on state i's AI capabilities.

A minor conflict is triggered if either state chooses non-zero attacks. If a conflict is triggered, then each side suffers the costs of war

$$L_i(y_i, y_j; K_i, K_j) = c_m + (y_i + y_j)\frac{y_i y_j}{(y_i + y_j)^2} = c_m + \frac{y_i y_j}{(y_i + y_j)}$$

where $c_m > 0$ is the baseline cost of conflict, costs scale up the more powerful capabilities are deployed, and costs decrease the wider the gulf between states' demonstrated capabilities.

This model assumes that an attack of magnitude y_i = K_n is the only way to signal capabilities >= K_n?
I would expect that many capabilities thresholds could be signaled with an attack that's less than full powered (perhaps with no attack at all).

Date 3 (major conflict).

Each i chooses $Z_i \in \{0, 1\}$, where 1 means to attack and 0 means not to attack. If $Z_1 = Z_2 = 0$ (peace), the states split a prize V – $(V/2, V/2)$. If $Z_1 = 1$ or $Z_2 = 1$ (war):

1. If some state i has a decisive strategic advantage, then that state wins the entire prize V; the loser receives 0.

2. If neither state has a decisive strategic advantage, war is catastrophic for both: each side loses θ.

We begin with both states having common knowledge of the parameters w, s_i, T, c_m, V, and θ.

4.2 Solving the model

We solve for a Perfect Bayesian Equilibrium by backward induction.[16]

At date 2, $y_j \in [0, K_j]$ is public knowledge and creates a credible lower bound on K_j. Given y_j, the opponent's posterior for K_j is the prior truncated to $[y_j, \infty)$, i.e. Let the probability mass cut off by observing y_i be:

$$B_j \ := \ \Phi(\frac{L_j - w}{\sigma}) \in [0, 1).$$

Date 3 (major conflict).

If war occurs, player i's expected payoff given private K_i and public y_j is

$$\mathrm{U}_i(K_i \mid y_j) = p_i(K_i \mid y_j)V + (1 - p_i(K_i \mid y_j))(-\theta),$$

where

$$p_i(K_i \mid y_j) = \Pr\big(K_i \geq K_j + T \mid K_j \geq y_j\big) = \begin{cases} 0, & K_i - T \leq y_j, \\ \dfrac{\Phi(\frac{K_i - T - w}{\sigma}) - B_j}{1 - B_j}, & K_i - T > y_j. \end{cases}$$

[16] Φ, φ denote the standard normal CDF and pdf respectively.

If peace occurs, each state receives $V/2$. Thus i attacks iff $U_i \geq V/2$, i.e.

$$p_i(K_i \mid y_j) \geq \tau \qquad \tau := \frac{V/2 + \theta}{V + \theta} \in (\tfrac{1}{2}, 1).$$

Solving $p_i = \tau$ yields a unique cutoff rule:

$$k_i^\star(y_j) = w + T + \sigma\Phi^{-1}(B_j + (1 - B_j)\,\tau)$$

$$K_i \geq k_i^\star(y_j).$$

$k_i^\star$ is increasing in y_j and T and θ, and decreasing in V. The stronger my opponent *appears*, the harder it is to achieve a DSA (larger T), or the more devastating it is to fight without a DSA (larger θ), the more selective I become about attacking.

Date 2 (minor conflict).

Each i chooses $y_i \in [0, K_i]$; if either state chooses to attack, then each pays

$$L_i(y_i, y_j) = c_m + \frac{y_i y_j}{y_i + y_j}.$$

Choosing y_i has an informational effect: it shifts the public lower bound, which shifts the opponent's attack cutoff $k_j^\star(y_i)$ upward. Two observations determine the equilibrium form.

1. **All or nothing.** Fix beliefs about y_j. If $y_j = 0$, then any $y_i > 0$ costs exactly c_m, but a larger y_i deters more by raising $k_j^\star(y_i)$. Hence, if your opponent holds off, then the best response is $y_i = K_i$. If $y_j > 0$, raising y_i both increases deterrence and the closeness cost $\frac{y_i y_j}{y_i + y_j}$; given the functional form of costs from minor conflict, the net gain from a small increase at interior y_i remains positive, so the best response is K_i.

2. **Only peace-preferring types choose a minor conflict.** If $K_i \geq k_i^\star(y_j)$, i will attack at Date 3. A minor conflict cannot change the outcome at the final date, because war occurs if either state attacks, and it incurs a cost. Hence only types with $K_i < k_i^\star(y_j)$ might engage in a minor conflict to *deter* the opponent.

Define the opponent's Date 3 attack probability when you reveal a lower bound y_i:

$$A(y_i) \ := \ \Pr(K_j \geq k_j^\star(y)) = 1 - \Phi(\frac{k_j^\star(y_i) - w}{\sigma}).$$

Revealing $y_i = K_i$ at date 2 reduces the attack probability by

$$\Delta A(K_i) \ := \ A(0) \ - \ A(K_i),$$

which is strictly increasing in K_i because $k_j^\star(y_i)$ is increasing in y_i.

Let $\text{Harm}_i(K_i)$ denote the harm you avoid by deterring your opponent from attacking at Date 3:

$$\text{Harm}_i(K) \ := \ (\frac{V}{2}) - E[U_i | Z_j = 1].$$

In the catastrophic regime, where θ is so large that an opponent attacks only when doing so is almost surely decisive, $\text{U}_i \approx 0$ and $\text{Harm}_i(K_i) \approx V/2$ for all peace-preferring states. Using this approximation simplifies the algebra below.

The expected cost of engaging in a minor conflict with full capability revelation is

$$\text{L}_i(K_i) \ = c_m + E\left[\frac{K\, K_j}{K + K_j}\right].$$

Date 2 best response. A state that prefers peace at Date 3 engages signals through a minor conflict if, and only if

$$\Delta A(K_i) \cdot \text{Harm}_i(K_i) \geq L_i(K_i).$$

Substituting our approximation, we obtain:

$$\Delta A(K_i) \cdot \frac{V}{2} \geq L_i(K_i).$$

Because $\Delta A(\cdot)$ and $\text{Harm}_i(\cdot)$ are (weakly) increasing and the closeness cost is (weakly) increasing in K_i, the left-hand side crosses the right-hand side once. Hence, there is a (weakly) unique cutoff $\hat{K}_i \in [0, k_i^\star(0)]$ satisfying

$$\Delta A(\hat{K}_i) \cdot \frac{V}{2} = L_i(\hat{K}_i),$$

such that

$$y_i(K_i) = \begin{cases} 0, & K_i < \hat{K}_i, \\ K_i, & \hat{K}_i \le K_i < k_i^\star(0), \\ 0, & K_i \ge k_i^\star(0). \end{cases}$$

The upper bound $k_i^\star(0)$ is the capability level past which there are worlds where state i will attack no matter what – namely those worlds where $y_j = 0$. **Perfect Bayesian Equilibrium.** By symmetry ($w_1 = w_2 = w$) and in the catastrophic regime ($\text{Harm}_i(K_i) \approx V/2$), the equilibrium is:

$$\text{Date 2:} \quad y_i(K_i) = \begin{cases} 0, & K_i < \hat{K}_i, \\ K_i, & \hat{K}_i \le K_i < k^\star(0), \\ 0, & K_i \ge k_i^\star(0), \end{cases}$$

$$\text{Date 3:} \quad Z_i = \begin{cases} 1, & K_i \ge k_i^\star(y_j), \\ 0, & K_i < k_i^\star(y_j), \end{cases}$$

Comparative statics.

1. If T increases or θ increases, and so $k_i^\star(\cdot)$ increases; fewer attacks occur in expectation; the gain from deterrence $\Delta A(K_i)$ rises, so $\hat{K}$ weakly falls (such that more states engage in minor conflict).

2. If V increases, then τ decreases and Harm increases; states with powerful capabilities attack more readily *and* states with intermediate capabilities engage in more minor conflict.

3. If c_m increases, then $\hat{K}$ increases: there is less minor conflict and more major conflict in expectation.

4. If σ increases, then [TODO].

In this model, the final war decision has a clean cutoff: given your opponent's revealed lower capabilities bound, you attack if, and only if, your private capabilities exceed some threshold. In anticipation of this cutoff, the minor conflict stage exhibits a stark pattern. Only states with intermediate AI capabilities, who would not attack without further information, choose to initiate a minor conflict. When doing so, they go all-in to maximize deterrence. This result is interesting: it predicts selective transparency and under-revelation emerging from the *costs* of a minor conflict.

The knife-edge features, however, are fragile. Changing the functional form of the costs of minor conflict can yield interior solutions at Date 2 rather than all or nothing results. In addition, allowing non-catastrophic major conflicts in the absence of a DSA may induce more non-decisive attacks. Despite these aspects of fragility, the core logic of the model is robust: attacks would remain cutoff-based and minor conflict signaling would remain concentrated among states with intermediate levels of AI capabilities.

5 Extending the Model

This model is simple in concept – it is aimed primarily at demonstrating the importance of uncertainty when it comes to the strategy of conflict when a decisive strategic advantage seems to be at stake. We lay out several promising avenues for extending the model below; the list is not exhaustive.

1. **Endogenize capabilities.** States can choose how much to invest in AI before uncertainty resolves, so they can make strategic trade-offs. In the model, states could make a choice at Date 1 that partially determines capabilities.

2. **Endogenize escalation.** Minor clashes can slip into major war: per Schelling, the creation of risk can be a bargaining chip. In the model, we could introduce another choice of a hazard level h_i, and apply it to the cost function.

3. **Endogenize information design.** States can choose to disclose information with different statistical profiles, not only via lower-bound-setting conflict. In the model, we could add a pure signaling stage, or allow states to choose a more nuanced signaling profile at Date 2.

4. **Asymmetry.** Real rivals need not be symmetrical. Asymmetry could, in principle, significantly alter decisions made at each stage of the game. A particularly interesting extension may be combining asymmetry with endogenous escalation in order to explore some the dynamics in Hendrycks et al. (2025).

5. **Repeat interactions and time evolution.** AI races unfold over time: capabilities evolve, learning accumulates, and reputations form. This evolution may create path dependence and windows of vulnerability or opportunity. While more elegant time-continuous models may be possible, a model where capabilities accumulate over time and the basic model repeats indefinitely (or until DSA-enabled victory or catastrophe) could be a useful starting point.

6 Policy Recommendations

[TODO]

Conclusion

[TODO]

References

David Abecassis. Refining MAIM: Identifying Changes Required to Meet Conditions for Deterrence. 2025.

Dario Amodei. Machines of Loving Grace. 2024.

Stuart Armstrong, Nick Bostrom, and Carl Shulman. Racing to the Precipice: A Model of Artificial Intelligence Development. *AI and Society*, 31(2):201–206, 2016.

Leopold Aschenbrenner. Situational Awareness. 2024.

Peter Barnett and Aaron Scher. AI Governance to Avoid Extinction: The Strategic Landscape and Actionable Research Questions, 2025. URL `https://arxiv.org/abs/2505.04592`.

Nick Bostrom. *Superintelligence: Paths, Dangers, Strategies.* Oxford University Press, 2014.

B. Boudreaux, G. Smith, E. Geist, and L. Dion. Insights from Nuclear History for AI Governance. 2025.

Bernard Brodie and Frederick Sherwood Dunn, editors. *The Absolute Weapon: Atomic Power and World Order*. Harcourt, Brace, New York, 1946.

James Coates. AI and Nuclear Deterrence. forthcoming.

Nicholas Emery-Xu, Andrew Park, and Robert Trager. Uncertainty, Information, and Risk in International Technology Races. *Journal of Conflict Resolution*, 68(10):2019–2047, 2024.

James D. Fearon. Domestic Political Audiences and the Escalation of International Disputes. 1958.

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

James D. Fearon. Signaling Foreign Policy Interests: Tying Hands versus Sinking Costs. *The Journal of Conflict Resolution*, 41(1):68–90, 1997.

Dan Hendrycks, Eric Schmidt, and Alexandr Wang. Superintelligence Strategy: Expert Version, 2025. URL https://arxiv.org/abs/2503.05628.

Herman Kahn. The Nature and Feasibility of War and Deterrence. 1960.

Emir Kamenica and Matthew Gentzkow. Bayesian Persuasion. *American Economic Review*, 101(6):2590–2615, October 2011.

Corin Katzke. Is a Larger US Lead in AI Development Valuable? unpublished.

Corin Katzke and Gideon Futerman. The Manhattan Trap: Why a Race to Artificial Superintelligence is Self-Defeating, 2024. URL https://arxiv.org/abs/2501.14749.

Robert M. McNamara. Congresional Record. 1967.

J. J. Mearsheimer. *The Tragedy of Great Power Politics, publisher = Norton Company*. 2001.

H. J. Morgenthau. *Politics among nations: the struggle for power and peace*. Alfred A. Knopf, 2001.

L. C. R. Patell. Strategic Transparency. forthcoming a.

L. C. R. Patell. The Dangers of MAIM. forthcoming b.

L. C. R. Patell and O. Guest. Signaling Restraint. forthcoming a.

L. C. R. Patell and O. Guest. A U.S. AI Project is Not the Same as Racing. forthcoming b.

Iskander Rehman, Karl P. Mueller, and Michael J. Mazarr. Seeking Stability in the Competition for AI Advantage. 2025.

T. C. Schelling. Prospectus for a Reorientation of Game Theory. 1958.

T. C. Schelling. *The Strategy of Conflict*. Harvard University Press, 1960.

U.S. Department of Defense. Annual Review of Defense Programs. 1966.

Kenneth N. Waltz. *Theory of International politics*. Addison-Wesley Pub. Co., 1979.

Peter Wildeford and Oscar Delaney. Mutual Sabotage of AI Probably Won't Work. 2025.

Sam Winter-Levy and Nikita Lalwani. The End of Mutual Assured Destruction? 2025.

Albert Wohlstetter. The Delicate Balance of Terror. 1958.