TOBIAS SYTSMA, ALVIN MOON

Can AGI Rivals Hold Back?

Verification, Domestic Oversight, and the Conditions for Mutual Restraint

This research was independently initiated and made possible by RAND's own resources, including income from operations, unrestricted gifts from RAND supporters, and distributions from its long-term investment fund.

Research Report

For more information on this publication, visit **www.rand.org/t/RRA5163-1**.

About RAND

RAND is a research organization that develops solutions to public policy challenges to help make communities throughout the world safer and more secure, healthier and more prosperous. RAND is nonprofit, nonpartisan, and committed to the public interest. To learn more about RAND, visit www.rand.org.

Research Integrity

Our research integrity is grounded in RAND's core values of quality and objectivity. Rigorous quality assurance procedures, conflict of interest screening, and transparency in funding ensure that every study is objective and nonpartisan. Learn more at www.rand.org/integrity.

RAND's publications do not necessarily reflect the opinions of its research clients and sponsors.

Published by the RAND Corporation, Santa Monica, Calif.
© 2026 RAND Corporation
RAND® is a registered trademark.

Limited Print and Electronic Distribution Rights

This publication and trademark(s) contained herein are protected by law. This representation of RAND intellectual property is provided for noncommercial use only. Unauthorized posting of this publication online is prohibited; linking directly to its webpage on rand.org is encouraged. Permission is required from RAND to reproduce, or reuse in another form, any of its research products for commercial purposes. For information on reprint and reuse permissions, visit www.rand.org/about/publishing/permissions.

RR-A5163-1

About This Report

Under what conditions can rival states make mutual restraint in their development of artificial intelligence (AI) credible and self-enforcing when neither can directly observe the other's compliance? We place this question in the context of a game-theoretic model to identify the strategic requirements for mutual restraint. In the model, each state first chooses how much oversight to impose on its own AI industry and then decides whether to comply, abandon, or covertly defect from a restraint arrangement. This type of analysis enables us not only to identify the conditions under which restraint is feasible and sustainable but also to show how these conditions shift as states come to believe that they are close to attaining artificial general intelligence (AGI). The analysis also explains why rivals may fail to build the necessary domestic oversight capacity even when both rivals would prefer to restrain AI development. The report should be of interest to policymakers and analysts responsible for U.S. strategy on advanced AI and for researchers studying AI governance.

Center for the Geopolitics of Artificial General Intelligence

RAND Global and Emerging Risks is a division of RAND that delivers rigorous and objective public policy research on the most consequential challenges to civilization and global security. This work was undertaken by the division's Center for the Geopolitics of AGI, which is committed to helping decisionmakers understand, anticipate, and prepare to navigate the national security and geopolitical implications of AGI. The center convenes leading technologists, strategists, economists, political scientists, and outside experts to consider the feasibility and effectiveness of prospective AGI-enabled capabilities; the domestic and international implications of their use; and the strategies and policies that governments, businesses, and civil society could adopt to respond to new realities. For more information, visit www.rand.org/geopolitics-of-agi.

Funding

This research was independently initiated and conducted within the Center for the Geopolitics of AGI using income from operations and gifts from RAND supporters, including philanthropic gifts made or recommended by DALHAP Investments Ltd., Ergo Impact, Founders Pledge, Charlottes och Fredriks Stiftelse, Good Ventures, Longview, and Coefficient Giving. RAND donors and grantors have no influence over research findings or recommendations.

Acknowledgments

This report benefited from conversations with RAND colleagues, such as those working in RAND's Center for the Geopolitics of AGI and attendees at the Post-AGI Civilizational Equilibria workshop held in May 2026. In particular, we would like to thank Joel Predd, Benjamin Boudreaux, Matt Chessen, and Hal Brands for their input and guidance. Thank you to Scott Behmer and Jane Vaynman who served as peer reviewers. We used Claude Fable 5 and ChatGPT 5.6 to explore ideas, formalize and solve the game-theoretic model, draft and refine prose, and review and critique the report. All remaining errors are our own.

Contents

About This Report iii
Tables vii

Can AGI Rivals Hold Back? Verification, Domestic Oversight, and the Conditions for Mutual Restraint 1
Introduction 1
The Model in Brief 3
What Self-Enforcing Restraint Requires 5
Why Restraint May Still Fail 13
Strategic Implications and Limitations 15
Limitations 17
Conclusion: A Strategic Test 18

APPENDIX
Technical Discussion of the Formal Model 21

References 35
About the Authors 37

Tables

1. Classes of Defection and What Deters Each 6
A.1. Table of Assumptions 23
A.2. Example Defections and Their Relationships to the Model 29

Can AGI Rivals Hold Back? Verification, Domestic Oversight, and the Conditions for Mutual Restraint

Introduction

The question of whether, and how, to slow down the pace of artificial intelligence (AI) development now sits near the center of the policy debate. However, any effort to rein in AI development must contend with strong underlying competitive pressures. OpenAI's CEO, Sam Altman, has argued that the United States must maintain its lead in AI development as authoritarian governments seek to catch up (Altman, 2024). Similarly, Dario Amodei, the CEO of Anthropic, states that slowing U.S. AI progress is untenable, in part because authoritarian countries would continue their push to develop the technology even if democratic countries slowed their efforts (Amodei, 2026). At the same time, more than 1,000 employees of frontier AI companies—including senior researchers and executives from OpenAI, Anthropic, Meta, and Google DeepMind—have signed "Pacing the Frontier," a July 2026 open letter requesting that the U.S. government support an international effort to pace certain aspects of AI development.

The competing concerns over competition and cooperative restraint illustrate a familiar strategic dilemma. Each actor might believe that it would be better if everyone would hold back from engaging in the most destabilizing forms of competition, but each might be uncertain about whether others will actually restrain their AI development efforts or be tempted to push ahead and capture a first-mover advantage. If the temptation of achieving a first-mover advantage is strong enough, or trust among actors is weak enough, the result of this dilemma could be a race that no actor wants.

Prior work conducted in RAND's Center for the Geopolitics of Artificial General Intelligence (AGI) shows that a two-country AGI competition can be a prisoner's dilemma when the advantage of reaching AGI first is large relative to the shared risk of accelerated development and that the competition becomes a coordination game once that shared risk grows large enough (Abraham, Kavner, and Moon, 2025). Root (2026) reaches a similar conclusion, suggesting that cumulative catastrophic risk can make restraint rational, even without any formal agreement between parties. The Freedom of Action framework presented in Predd et al. (2026) treats the feasibility of restraint among rivals as one of five pivotal uncertainties that AGI strategy must account for, but the authors leave open what rivals must do to make restraint credible.

This analysis begins where that prior RAND research left off. Specifically, we developed a game-theoretic model in which two rivals can prefer to exercise mutual restraint in AI development; however, they fail to sustain this restraint because neither can verify that the other is holding back. We add the dimension that some of the information required for verification may only be generated when states have built the domestic institutions capable of collecting and preserving evidence of their compliance. Thus, a state that has never made its AI industry legible, even to itself, cannot show that it is complying, even when it is. Therefore, whether restraint is feasible depends, in part, on what each state has built at home to make its *own* conduct verifiable. In the two-stage model that we developed, whether restraint between rivals is feasible depends on the ways in which a state could cheat, on how detectable each kind of cheating is, on the cost (financial, political, or otherwise) of making a state's domestic AI ecosystem legible to a rival, on the expected size of the strategic advantage that abandoning cooperation would create, and on how much time remains before AGI arrives.

The problem of verification in the AI context may differ from the classic arms control formulation of the problem. The premise in the traditional arms control context is that a state generally knows what is occurring within its own programs. This makes verification primarily a question of how much of that information the state is willing to expose to a rival. In frontier AI, development activity is dispersed across private labs that are

also operating in a highly competitive environment and in a new industry that does not have an established mechanism for collecting, processing, and centralizing the information that might be required for verification purposes. In such a setting, a state's capacity to know what is happening within its own AI ecosystem becomes a critical factor of the international verification equation.

In this report, we discuss several strands of research on cooperation under high-stakes competition with potentially catastrophic outcomes. Prior research has characterized when the AGI race is a prisoner's dilemma and when it can become a coordination problem (Abraham, Kavner, and Moon, 2025) and has shown that transparency and trust can shape whether rivals can coordinate to pause AI development (Fudenberg and Koh, 2026). Heijnen and Dam (2019) addresses international cooperation under catastrophic risk in the context of climate change, showing that when a catastrophe can irreversibly change the nature of the game that rivals are playing, cooperation may not follow the logic of standard repeated games because an irreversible end to a cooperative relationship can weaken the rewards and punishments that usually make cooperation sustainable. The arms control literature also suggests that restraint is feasible only if compliance can be verified. Coe and Vaynman (2020) explains why arms control has been historically rare, showing that states face a trade-off between providing enough transparency to assure their compliance with an arms control arrangement without exposing enough information to threaten that state's national security. Baker (2023) documents how verification operates in nuclear practice, where it rests on states' self-declarations of facilities and materials, checked by inspection, with national technical means of verification serving mainly to flag what a state fails to declare. We build on these studies by treating the conditions that make restraint possible as a choice that states make in the context of the strategic environment.

Key Takeaways

- **Domestic AI oversight can enable international restraint commitments.** Verification that a state is complying with a restraint commitment depends, in part, on domestic oversight because a state opaque to itself may not be able to prove its compliance, even when it is complying. Thus, making one's own potential cheating detectable can make restraint more credible to a rival.
- **Fear of catastrophe has deterrent value.** Fear of a catastrophic transition to AGI deters some dangerous forms of cheating on a restraint commitment, while optimism about the transition to AGI raises the need for verification in an arrangement for mutual restraint.
- **Verification deters cheating when players believe that AGI is a distant possibility but loses traction as rivals come to believe that AGI is imminent.** If punishment for cheating on a restraint commitment is delivered before AGI arrives, believing that AGI is closer will leave less time to impose costs on the defector. As a result, punishment loses its force as AGI arrival nears, regardless of how good detection of violations becomes. Near AGI arrival, instruments tied to post-AGI outcomes, such as benefit-sharing or forfeitable stakes, may still deter cheating.
- **Even if restraint is mutually preferable, it may still fail when one rival doubts that the other will make its actions verifiable.** If domestic reasons for investment in oversight are sufficiently weak, a state will build the domestic oversight needed for credible verification only if it expects its rival will do so, so mutual doubt can foster an opaque world even when both states would prefer a legible world. Building the means for verification does not commit states to restrain their AI development, but it makes that choice available to them.

The Model in Brief

The model has two stages. In the first stage, each state chooses how much oversight to impose on its own AI industry. That choice is measured as the additional oversight a state adopts for the sake of international verification, on top of whatever governance it would maintain purely for domestic reasons.[1] Oversight might include such mechanisms as registries, licensing, compute accounting, evaluation records, and audit systems or, more generally, whether the information that inspectors require has been structured and governed so that it leaves an observable trace. We do not model the explicit form of domestic oversight that states choose, and the model does not impose a strict assumption that states must invest in oversight capacity, but we do assume that oversight capacity is costly to produce and hard to reverse once created. The level of domestic oversight that a state selects in the first stage of the game partially determines how detectable the state's own future cheating would be in the second stage of the game.

In the second stage of the game, the two states can decide to form an arrangement to slow down their AI development. In this stage, states interact period by period until AGI arrives at a random future date. The arrival of AGI ends the game, and we discuss this feature of the model later in this section. Each period before AGI arrives, a state either complies with restraint, races openly (i.e., makes it publicly known that it no longer wants to abide by the arrangement), or covertly defects.[2] Covert defections may differ in the immediate benefit they produce for the cheating state, in how much they improve the cheating state's odds of reaching AGI first, and in how much catastrophic risk they create. If a covert defection is detected by the other state, the arrangement collapses permanently, and both states fall back to racing toward AGI.

As noted earlier, the arrival of AGI terminates the game. This does not assume that AGI marks the end of political or economic history but does represent the boundary of the pre-AGI race examined in this analysis.[3] Once a state reaches AGI, the strategic question changes from whether rivals will restrain their efforts to develop the technology to how the resulting advantage and risks of the transition will be distributed, and the latter world is summarized in payoffs that players receive when the game ends, rather than modeled explicitly.

The strategic interactions between states are structured around two baseline assumptions about the nature of the technology. First, we assume that the state arriving at AGI first receives a higher payoff (a *first-mover advantage*) than the state that arrives second. Whether this advantage is large or small partially determines states' actions in the game, and the model imposes the assumption that such an advantage is positive and finite. Second, we assume that a bad transition to AGI imposes a common harm on both sides and that a transition that occurs when players are racing toward AGI is worse than a transition that occurs when players are mutually restrained. The model does not make any assumption about how much worse, in terms of economic loss or other loss, a transition is under the condition of both rivals racing toward AGI, only that

[1] Oversight of a state's own domestic AI ecosystem pays off internationally if it helps make the state's future cheating detectable, hence punishable, hence deterred, which makes the state's promise of restraint credible enough for a rival to rely on. This is Schelling's power to bind oneself (Schelling, 1960), implemented in regulatory infrastructure. In this interpretation, domestic AI governance can be seen as a foreign-policy instrument. The registries, licensing regimes, and audit systems that make an industry governable at home are the same entities that make the state's international commitments believable abroad. A state may therefore rationally regulate its own AI industry *more* strictly than its domestic interests alone would justify, and the burden of that regulation has strategic value that might be invisible to a purely domestic cost-benefit test.

[2] The public rejection of the restraint arrangement could come from several external factors, such as political regime changes or elections, which we do not model directly. However, to the extent that these changes result in the state letting it be known (implicitly or explicitly) that it no longer wants to partake in the arrangement and is going to race openly toward AGI, the effects are captured in the model. Public rejections of the arrangement differ, of course, from covert defections, in which that state continues to act publicly as if it is upholding the arrangement while secretly racing toward AGI.

[3] Our modeling approach follows that of Abraham, Kavner, and Moon (2025).

the magnitude of the damage can be ordered. These assumptions, and their implications for the model, are discussed below and in the appendix.[4]

Three results follow:

1. **Expected catastrophic harm can lower the amount of formal verification required to sustain restraint.** If a state expects to bear some of the risk that it creates through a defection, the defection is partly self-deterring. However, this self-deterrent effect is only as strong as the defector's perceived risk of a catastrophic transition to AGI and its belief that its defection increases that risk. At the same time, fear of a failed transition to AGI can make the cooperative restraint arrangement more valuable: If a player is caught cheating, racing resumes. The more optimistic the states are about the safety of the transition to AGI, the more verification is required to sustain the arrangement.
2. **Short timelines weaken enforcement unless punishment occurs after AGI's arrival.** In the model, the punishment for defection is the loss of the future benefit of cooperation. A state that cheats today loses the cooperative relationship tomorrow. As the expected arrival of AGI draws closer, the number of remaining tomorrows shrinks, weakening any punishment whose force lies in future consequences. Eventually, a sufficiently tempting defection cannot be deterred at all, even when detection is certain. If states expect AGI to arrive soon, restraint can only be held together by threats of punishment that are explicitly tied to outcomes *after* AGI arrives and conditioned on pre-AGI defections.
3. **States may fail to build sufficient detection capacity, even if both would benefit from it.** Mutual restraint requires each state to make its own potential cheating sufficiently detectable to the other side, but building the domestic oversight that creates this visibility is costly, both in financial and political terms. A state has less incentive to bear the cost of building domestic oversight, making its own future defections visible, if it expects its rival to remain opaque. Therefore, the states' oversight choices form a coordination game with two stable outcomes: a *legible world* in which both states build oversight capacity and restraint is possible or an *opaque world* in which neither builds sufficient oversight capacity and restraint is impossible. The opaque world can prevail even when the legible world is better for both states. Because restraint requires each side to make its AI ecosystem legible to the other, the state that faces a more costly effort to become legible can block whether a restraint arrangement is available to either side.

These results have implications for AGI strategy. First, they suggest that domestic AI oversight has value outside purely domestic regulatory objectives. The domestic institutions and capabilities that allow a state to gain visibility into its own AI ecosystem can also, under certain conditions, provide the foundation on which restraint arrangements can sit. The above results also provide a lens through which current proposals for AI governance can be viewed. Policy proposals—such as compute reporting thresholds (Bureau of Industry and Security, 2024), third-party audits (Brundage et al., 2026), incident reporting (U.S. House of Representatives, 2026), and on-site inspection—appear in the contemporary policy discourse (Larsen et al., 2026), and the model suggests that the value of these policies may include the act of creating the scaffolding that would enable data collection and inspection, as well as the purported benefits that the policy explicitly

[4] In the model, we hold political authority and the strategic environment fixed to focus on failure modes that arrive because of defection from the cooperative arrangement. Therefore, the model does not capture elections, leadership turnover, policy realignment, bureaucratic loss of control, or other changes that could cause a state to abandon restraint for reasons that are independent of the strategic incentives modeled here. The region of the model where restraint holds should consequently be understood as a result that is conditional on political continuity. The net effect of adding these changes is ambiguous because they may introduce ways in which a restraint arrangement can break down or they could strengthen self-deterrence and make cooperation easier to sustain. We defer this analysis to future work. For further discussion, see the "Limitations" section on page 17.

states. Lastly, because mutual restraint requires both states to be legible to each other, a state may sometimes be self-justified in helping a rival become more verifiable. As a result, if the value of a restraint arrangement is sufficiently high for the side where legibility is less costly, financing a rival's verification capacity can be a rational act. We discuss these implications, and several others, in the "Strategic Implications and Limitations" section on page 15.

What Self-Enforcing Restraint Requires

The model is solved backwards, meaning that the solution requires identifying the defections that a potential arrangement can deter first (Stage 2), then determining whether players will build the relevant domestic capabilities for that deterrence to occur (Stage 1). This section discusses the second stage of the game.

A Menu of Defections

We refer to the list of defections a state could choose from as a *menu of defections*. The menu represents the idea that racing toward AGI is not a single act. For instance, a state might cheat by running more compute than it has declared, by hiding a frontier facility, by weakening safety processes, by stealing a rival's model weights, by manipulating audit reports, by routing activity through a third country, or by covertly accumulating a software advantage that leaves little physical trace. These acts create different temptations and leave different evidence, and a verification system that catches one of them could miss the others entirely.

Regardless of which option a state chooses, each covert defection, denoted as d, has four relevant features. The first is the near-term benefit of cheating, which we denote as g_d. This term represents the value a state gets each period it cheats and is not caught. This near-term value can include such benefits as saved compliance costs and faster deployment timelines. The second is the positional gain, denoted as π_d, which captures how much the defection improves the state's chance of reaching AGI first, earning it a potential decisive advantage. The third is the added catastrophic risk, ρ_d, that the defection creates. Some defections may greatly increase the perceived risk of catastrophe, while others may shift the probability of a catastrophic transition to AGI by very little. The fourth is the probability of detection, $q_d(q_0, x_i)$, which depends on the oversight infrastructure (x_i) that the cheating state has built into its own AI ecosystem (an equilibrium value determined in Stage 1) and a baseline level of detection (q_0) that can be achieved unilaterally, regardless of the oversight infrastructure. For the purposes of this analysis, we hold that baseline level fixed to isolate the implication of the oversight infrastructure that the cheating state has built.

These four features can help address the following relevant policy questions: How attractive is the defection (which operates through g_d and π_d)? How much does the defector fear the risk it creates (ρ_d)? How observable would the act be if the state tried to hide it (q_d)? The answers differ across the menu of defections, and these differences drive what follows from the model. Table 1 organizes the menu of defections into classes, by what deters each one, and Table A.2 in the appendix links these example defections to their relevant model parameters.

Some Defections May Be Particularly Difficult to Detect

Different forms of AI development and racing produce different types of evidence. For instance, compute-heavy cheating has a physical footprint because large training runs require chips, data center capacity, power, cooling, procurement, and personnel, and each of these components leaves traces that registries and audits can surface. Baker (2023) draws an analogy to the nuclear setting, noting that verification regimes begin with state declarations, domestic accountancy, and inspections that test whether declared facilities and materials match the record. Baker et al. (2025) breaks AI verification into claims about declared large-scale compute

TABLE 1
Classes of Defection and What Deters Each

Class	Example Forms	Primary Source of Temptation	What Deters Defection in the Model
Risk-creating (potentially self-policing)	• Skipping safety evaluations • Weakening safeguards • Rushing deployment	Mostly speed and cost savings, less positional gain	The defector's own share of the added catastrophic risk, which may require no detection and no rival response if self-policing effect is strong enough.
Positional, visible to oversight	• Unauthorized large training runs • Undeclared frontier facilities	Improved odds of reaching AGI first	Detection: The cheater's own registries and audits can surface defections, and the intelligence floor may already be high.
Positional, hard to detect	• Exfiltration of a rival's model weights • Secret algorithmic gains • Internal automated R&D	Improved odds of reaching AGI first	Detection, but neither unilateral collection nor feasible cooperative oversight may detect it. This class of defection may be particularly difficult to detect.
Overt	• Open breakout from the arrangement	Maximal speed and position	Detection: The defecting state makes it clear to its rival that it no longer wishes to remain bound by the arrangement.

NOTE: The classes are ideal types in the context of the model, and real actions can overlap them. A covert training run is both positional and risk-creating if it pushes a frontier model past safety controls. A safety shortcut becomes positional if it lets a state cross a deployment threshold first. The model does not need every real-world defection to fit one row. It needs the menu to be explicit enough so that analysts can estimate, for each item, the benefit, the position it buys, the risk it adds, and the evidence it would leave. R&D = research and development.

use and claims that no undeclared large-scale use exists. Because compute is specialized and physically large, at least relative to other AI inputs (such as data and algorithms), compute accounting may be less intrusive than trying to account for every input to AI progress.

Other inputs to AI development, including algorithms, data, model weights, and organizational know-how, have smaller physical footprints. Algorithmic progress alone has historically halved the compute needed to reach a given capability level roughly every eight months (Ho et al., 2024), but it is likely more difficult to observe such progress than the large training runs it displaces because algorithmic gains lack the procurement and energy signature that compute-intensive work leaves behind. A restraint regime built mainly around compute accounting may catch hidden training runs while missing exfiltrated weights, secret algorithmic gains, or internal automation that raises what a declared cluster can produce.

The Hardest-to-Deter Defection Sets the Burden of Verification

A rational cheater routes its cheating through whichever channel is hardest to monitor, conditional on having the resources to access that channel. Therefore, the arrangement's verification burden is set by the hardest-to-deter item on the menu of defections, in the same way that a verification regime is only as strong as its weakest layer (Baker et al., 2025).[5] In the model, each defection has its own required detection probability, so the cooperative arrangement is operational only if oversight meets the highest of these requirements.

[5] The claim here is not that narrow restraint arrangements are impossible. States could write a narrow arrangement that focuses specifically on covering one class of conduct and sustaining restraint in one specific dimension (e.g., a moratorium on recursive self-improvement specifically), while leaving other dimensions of competition uncovered by the arrangement. Even in this narrow case, there are defections that the arrangement must cover, and the principle still holds that the hardest-to-deter defection will set the verification burden. Narrowing the menu of defections lowers the verification burden if it removes hard-to-detect defections; however, doing so leaves channels open for defection to occur. If those channels provide substitutes for the conduct being restrained, such substitutes may reduce the arrangement's practical value, even if the arrangement succeeds on its own terms. Whether the resulting partial restraint remains worth sustaining is a policy judgment that the baseline model does not address.

Thus, the idea of a menu of defections has a strategic implication. Because there are many forms of cheating with different observable signatures, one productive question to ask of a verification and restraint arrangement is which row of Table 1 determines how demanding verification must be for restraint to hold. If the binding defections are hidden training runs, compute accounting may be sufficient for deterring most defections. If the binding defections are weight theft or secret algorithmic progress, compute accounting is still valuable but cannot carry the arrangement, a point we return to in the "Strategic Implications and Limitations" section on page 15.[6]

Verification as a Cooperative Technology

The institutional premise of this analysis is that domestic systems can expand what external monitoring is able to verify. While unilateral intelligence collection can reveal some cheating without any cooperation, and international inspectors can generate observations of their own, verification regimes work, in part, because the suspected party has created records and accounting systems that inspection can test against. The model captures this premise with a simple detection function:

$$q_d(x) = q_d^0 + \kappa_d x. \tag{1}$$

Recall that x is the oversight and regulatory infrastructure that a state builds into its own AI industry, beyond what it would maintain for purely domestic reasons.[7] The detection floor q_d^0 is the detection probability that unilateral intelligence collection provides, regardless of the rival's institutions. The increment $\kappa_d x$ is the added detectability created by the cheating state's own oversight infrastructure, where κ_d measures how exposed defection d is to that infrastructure. The subscript d indicates the type of defection; Table 1 suggests that different defections expose themselves through different evidence. We walk through each term of this function in greater detail below.

The Intelligence Floor

The intelligence floor q_d^0 includes national technical means of verification, cyber collection, human and signals intelligence, market traces, public reporting, and victim-side security systems. Effectively, everything that does *not* require the monitored party's cooperation falls under this term. Baker et al. (2025) identifies national intelligence as one of six layers available for AI verification and notes that it is the layer that can feasibly be used unilaterally. The intelligence floor may be uneven across defections, being comparatively high for acts with physical signatures and low for acts whose evidence lives in software or is otherwise difficult to identify through intelligence alone.

The Oversight Increment

The oversight increment $\kappa_d x_i$ is how much additional detectability the defector creates by building domestic governance through such policies as compute accounting, licensing, facility registries, chip-level logging, safety-case audits, and incident reporting. These systems raise the probability that cheating will be discovered because they create records that can be cross-checked against each other and against outside evidence. This has a direct precedent in the nuclear domain, where declarations make it demanding to defect, because

[6] An illustration comes from the AI Futures Project's Plan A proposal for a U.S.-China AI deal, whose verification architecture is built almost entirely around compute declarations and data center monitoring. Even in the Plan A scenario's own successful telling, a rival intelligence service exfiltrates frontier model weights without being detected (Larsen et al., 2026).

[7] The linear equation is illustrative. The formal results only require that detection probability increase with oversight capacity.

the defecting state must maintain a larger and more consistent lie across registries, inspections, and accounts (Baker, 2023).

From a strategic perspective, this increment means that choosing oversight can make one's own future cheating more detectable to a rival. That is why domestic AI governance can function as international commitment infrastructure. The same systems that let a state govern its own frontier AI systems also let the state demonstrate its restraint abroad because these systems allow the state to credibly claim that it could provide adequate transparency in the first place.

The Transparency-Security Trade-Off

Domestic oversight capacity is costly to build. Money and bureaucracy are part of that cost, but a potentially larger part of that cost comes from the potential exposure of sensitive information to a rival. Monitoring that is intrusive enough to prove compliance with a restraint arrangement might also create espionage channels and targeting information for a rival.[8] Coe and Vaynman (2020) argues that this tension is why arms control agreements have been rare historically, and Vaynman and Volpe (2023, 2025) extend this logic to AI directly, arguing that AI's dual-use nature places it in a potential arms control "dead zone" where verification is challenging, if not impossible, without exposing critical national security secrets.

The information that matters most for AI verification overlaps heavily with the information that firms and states most need to protect, since weights and training code are critical to both commercial and strategic advantage. The model compresses all of these burdens—monetary, security, and political—into a single cost of oversight, denoted as $c(x_i)$. Technologies that allow states to demonstrate their compliance without disclosing any underlying secrets, such as the confidential-computing and trusted-hardware methods surveyed in Baker et al. (2025), can operate in the model by lowering $c(x)$. By lowering the transparency cost, these technologies change whether the oversight capacity required by an arrangement would be something states are willing to build. This is consistent with Coe and Vaynman (2020), in which the feasibility of arms control depends on the trade-off between the transparency that an arrangement requires and the security cost of providing that transparency to a rival.

The Value of Restraint

Detection only matters if being caught results in the cheater losing something valuable. In this model, once cheating is detected or a state races openly, the arrangement collapses, and both states revert permanently to racing.[9] The value at stake is the cooperative surplus, denoted as S, which represents the difference between the value of living in a world of mutual restraint and the value of living in a world of mutual racing. If S is large, being caught cheating destroys something valuable, and there is real punishment to deter a potential cheater. If S is small, detection might reveal cheating without deterring it.

[8] This is an old tension within verification arrangements. Postwar proposals for international control of atomic energy collapsed when the Soviet Union would not accept inspection arrangements that would open its closed system to outside scrutiny, and secrecy proved a persistent obstacle to every control scheme on the table (Zaidi and Dafoe, 2021). Any AI restraint arrangement would likely face the same tension, since the political cost of legibility is not the same across political systems. We return to this in the "Unequal Costs of Legibility" subsection on page 14, where we show how the model determines which state's political constraints set the effective limit on oversight.

[9] In principle, other punishments for defection exist. States routinely link cooperation on one issue to conduct on another, and an AI restraint arrangement could be backed by export controls, tariffs, or other suspensions of unrelated cooperation. Punishments of this kind that operate before AGI arrives can be represented in the surplus because they destroy the pre-arrival value of the relationship. Punishments that would still be delivered, and meaningful, after AGI arrival lie outside the baseline model. We explain this distinction in the "Time and the Enforcement Horizon" subsection on page 12.

The surplus has a per-period component and a risk component. The per-period component is what restraint protects each period while restraint continues. The risk component is the probability of a safer transition that restraint preserves for whenever AGI arrives. Formally, this can be expressed as follows:

$$S = \frac{\Delta w + hL\Delta r}{1 - \beta}. \tag{2}$$

The term Δw is the ongoing value of cooperation (the *flow* value). This represents the ordinary gains from avoiding an open race, including lower security spending, less sabotage, fewer crisis-management costs, preserved trade in inputs, and shared safety research. The second part of the numerator is the risk value. Here, h is the perceived probability that AGI arrives in the current period, L is the harm each state expects from a catastrophic transition to AGI, and Δr is how much more likely catastrophe is when AGI arrival happens amid racing rather than amid restraint.

The severity parameter L represents an important belief that players hold in the model. A state convinced that the transition to AGI is survivable, or otherwise trivial, has a low L value. That optimism shrinks the value of the restraint arrangement itself. In the baseline version of the model, we treat a mismanaged transition to AGI as a common danger that both states expect to face, although the model leaves the magnitude of the disruption ambiguous. States may have different expectations about the dangers that AGI poses, and we address the implication of these differing beliefs later in this report. The model also treats a mismanaged transition as being more likely if AGI arrives when either state is racing toward its development. This modeling tradition follows prior RAND research on the game theory of AGI development (Abraham, Kavner, and Moon, 2025; Root, 2026). The denominator converts the per-period value into the value of an ongoing relationship, where β is the effective weight on the future.[10]

The flow and the risk components of the cooperative surplus behave differently. Flow value is present as long as the relationship continues and does not depend on states believing that AGI is likely to emerge soon. On the other hand, the risk value is larger when AGI arrival is perceived to be closer because the terminal consequences of the transition carry more weight inside each period. The model gives both states the same beliefs about proximity and severity, so AGI arrival being perceived as close applies to both players symmetrically. We discuss divergent beliefs later in this report. Proximity to AGI also shortens the expected life of the relationship, which shrinks the denominator's reach into the future. Proximity therefore raises the per-period stakes while cutting the horizon over which punishment operates. We return to this idea in the "Time and the Enforcement Horizon" subsection on page 12.

What the Surplus Does and Does Not Determine

A large cooperative surplus raises what a state would pay for oversight and determines whether the restraint is sustainable and how robust it is to doubt about a rival's actions. However, a large surplus, by itself, does not solve the detection problem that the states face. A relationship can be worth a great deal and still fail to hold the states to mutual restraint, either because the binding defections are nearly impossible to detect or because a defection's value is extraordinarily large. A state tempted by a positional defection (one that primarily improves its odds of getting to AGI first) weighs the expected first-mover advantage against the expected cost of being caught, and that expected cost is a product of detectability multiplied by the weight on the future multiplied by the surplus. If detectability is near zero, the other factors are irrelevant.

[10] The term β is a composite of both a standard discount rate δ and the per-period hazard rate h: $\beta = \delta(1 - h)$.

Two Empirical Assumptions

We make two explicit empirical assumptions in the model, and both are substantive claims about the world in which the players operate. First, we assume that $\Delta w > 0$: Mutual restraint protects the real flow value before AGI arrives. If that is wrong, the pre-arrival relationship is not worth very much, and enforcement far from arrival has almost nothing to protect.

Second, we assume that $\Delta r > 0$: A racing world is more accident-prone at AGI arrival than a restrained one. Rushed development, skipped safety evaluations, or weakened internal controls might confer a competitive edge in a racing dynamic, although any of these may be more likely to produce a loss-of-control event over deployed systems or a destabilizing crisis when AGI arrives. If this assumption does not hold, the risk component of Equation 2 disappears, and so does the relationship between the fear of catastrophe and verification because getting caught no longer consigns both states to a more dangerous world.

These assumptions can be contested, and the model shows exactly what contesting them implies. For instance, consider a state that believes the principal danger from AGI lies in *who,* specifically, reaches AGI first rather than in how fast anyone gets there. In such a situation, a rival's victory is the catastrophe. In the model's terms, that state is asserting a small Δr, a low payoff from finishing second, and consequently a large spread between winning and losing the race. A small Δr reduces the surplus that restraint protects, and a large first-mover advantage raises the temptation to defect in ways that improves the odds of winning the race. Both the smaller surplus and racing incentives make a restraint arrangement harder to hold together. Thus, states that locate the danger in each other, rather than in the transition to AGI, should expect verification requirements at the high end of the feasibility range. We return to what differing political systems do to the cost side of the restraint arrangement in the "Unequal Costs of Legibility" subsection on page 14.

The Deterrence Condition

This section describes the condition that determines, for any particular defection, whether a state will comply. A state comparing compliance with covert defection d weighs the expected benefit it would gain and the expected punishment from getting caught. It complies with the arrangement when the following incentive compatibility constraint holds:

$$\underbrace{g_d}_{\text{Immediate gain}} + \overbrace{h\pi_d(A-B)}^{\text{Positional gain}} \leq \underbrace{h\rho_d L}_{\text{Self-deterrence}} + \overbrace{\beta q_d(x_i)S}^{\text{Enforcement}}. \tag{3}$$

Incentives are compatible with restraint if the left side of Equation 3, which captures the expected gains, is less than or equal to the right side, which captures the expected costs. Each of the four terms corresponds with a policy surface that can, at least in principle, be measured or moved, which is what makes the deterrence condition useful rather than just compact. The following subsections walk through each term in Equation 3 specifically.

The Temptation Side

The first term of Equation 3, g_d, is the immediate gain from cheating while preserving the appearance of cooperation. Such gains might include time saved by skipping an evaluation, money saved by weakening controls, and speed gained by running undeclared compute. This gain arrives whether or not states perceive the arrival of AGI as being close.

The second term, $h\pi_d(A-B)$, is the expected positional gain, meaning how much the defection affects the probability that the cheater will reach AGI first. The factor π_d captures how much defection improves the state's chance of being first if AGI arrives while the cheating is underway. The spread $A-B$ is the additional value of finishing first rather than second (the first-mover advantage), which is where beliefs about decisive

strategic advantage enter the model. The hazard h scales the whole term by the likelihood that AGI arrival happens in the current period and explains why a state's perceived proximity to AGI can be detrimental for the cooperative arrangement. Specifically, a defection that barely matters when AGI feels distant can become compelling when states believe that the race is about to end because the potential decisive strategic advantage may be close to being achieved.

The Deterrence Side

The right side of Equation 3 contains the two forces that can hold a state back. The first, $h\rho_d L$, is self-deterrence. The terms h and L have already been introduced, but the ρ_d term requires explanation. This term captures how much of the catastrophic risk created from defection d the cheater expects to experience. If a defection adds catastrophic risk, and the defector expects to share in the resulting harm, then the defection carries its own penalty, with magnitude $\rho_d L$. Here, the penalty is incurred whether or not anyone detects the defection. Thus, if the expected catastrophic harm is felt even without any detection, and if the catastrophic harm is sufficiently large, then defections are not rational even without any detection (i.e., compliance is self-policing). However, the self-deterrence result holds only if the defector expects danger. A state that expects to survive, control, or offload the risk that it creates when it cheats has less self-deterrence; thus, the more optimistic states are about the transition to AGI, the more verification is required for the restraint arrangement to hold.

A defection is *self-deterred* when the defector's own risk exposure already outweighs the temptation:

$$\underbrace{h\rho_d L}_{\text{expected catastrophic risk the defection adds}} \geq \underbrace{g_d + h\pi_d(A-B)}_{\text{expected benefit the defection would confer}}. \tag{4}$$

For these types of defections, detection may be unnecessary. One implication is that purely positional defections—those with ρ_d at or near zero—can never be deterred by the expectation of a catastrophe alone. However pessimistic both sides become about the transition to AGI, a defection that is believed to cause little danger and remain undetected incurs no penalty. Fears of catastrophe discipline positional cheating only through the enforcement term, where they are multiplied by the detection probability.

The second term on the right side of Equation 3, $\beta q_d(x_i)S$, captures the expected loss of the cooperative relationship itself, resulting from direct enforcement of the arrangement's rules. It is a product of three factors: β is the weight on the future, which shrinks as states perceive AGI as being closer; $q_d(x_i)$ is the probability of getting caught and depends on the cheater's own oversight capacity; and S is the value of what collapse destroys (the outcome of Equation 2). Because the terms are multiplied, weakness in any factor weakens the whole term. Late timing, low detectability, or a low surplus can each make punishment too small to deter a state from being tempted to defect, no matter how strong the other two factors are.

A defection is *enforcement-dependent* when the inequality runs the other way. For these types of defections, restraint requires catching cheaters, and the required detection probability follows directly from the deterrence condition:

$$\overline{q}_d = \max\left\{0, \frac{g_d + h\pi_d(A-B) - h\rho_d L}{\beta S}\right\}. \tag{5}$$

Equation 5 can be read as a ratio of net temptation to available punishment. Larger immediate gains (g_d), larger positional payoffs $(h\pi_d)$, and a bigger first-mover advantage $(A-B)$ raise the detection probability that an arrangement must achieve $(\overline{q}_d)$. Greater self-deterrence $(h\rho_d L)$, a larger surplus (S), and a longer remaining horizon before AGI arrives (β) lower the detection probability. If the required detection probability exceeds what any feasible oversight system can deliver, that defection cannot be deterred within the model, and an arrangement that covers it fails.

Time and the Enforcement Horizon

This section describes how the perceived proximity of AGI arrival influences the incentive compatibility condition (Equation 3). Time enters the model through the hazard rate h, which captures the perceived probability that AGI arrives during this period. The model treats h as a fixed feature of the environment being compared, so statements about *rising* proximity are comparisons across worlds, with different perceived timelines, rather than a literal trajectory within a single game. However, the mechanism remains that punishment for defections loses force as the remaining pre-AGI future shrinks.

Proximity to AGI Arrival Changes Deterrence

Perceived proximity to AGI influences the effectiveness of deterrence. First, if states believe that AGI is close, the value of payoffs tied to its arrival grows. The positional temptation $h\pi_d(A - B)$ also grows because an improved chance of being first matters more when a first-mover advantage may be gained soon. The self-deterrence effect, $h\rho_d L$, grows for the same reason.

Second, proximity shrinks the future or, more specifically, the remaining future before AGI arrives. The weight on the future is $\beta = \delta(1 - h)$, which falls as h rises, and the enforcement term, $\beta q_d(x_i)S$, falls with it. As a result, the value of the cooperative restraint arrangement declines as states perceive AGI as a near-term possibility. In the limit, the model suggests that any punishment tied to pre-AGI actions ceases to be an effective deterrent. Mechanically, as h approaches one and β approaches zero, the enforcement term vanishes from Equation 3, even if defections are identified with certainty by the other state. The formal version of this statement is shown in Proposition 3 in the appendix.

However, just because direct punishment loses force when AGI is near, states' perceived risk about a catastrophic transition to AGI may still discipline their actions. When states' perceived risk of the fallback world (i.e., open racing) is sufficiently high, proximity to AGI can strengthen self-enforcement because reverting to a racing world is perceived as more dangerous when AGI arrival is plausibly near. The implication is that verification capacity, and the enforcement it provides, is most valuable in environments where AGI is not yet believed to be imminent.

Assumptions About AGI Arrival

These results are conditional on what the model assumes about AGI arrival, so if those assumptions do not hold, the results can change. For instance, if arriving at AGI first delivers no durable advantage, the positional temptation that players face in the game would disappear, and the limit would be harmless because there would be nothing left worth cheating for. Similarly, if the transition presented no danger, the risk component of the surplus (Equation 2) would disappear, and any self-deterrent effect from the potential of a catastrophic transition would disappear with it. If AGI arrival were a gradual process rather than a discrete event, the sharp boundary that the model draws between pre- and post-AGI periods would blur into a shrinking enforcement horizon over which punishment can occur (rather than vanish). We revisit these cases later in this report.

The model's *payoff structure* (i.e., what each player receives for their play) is also assumed to be history-independent, meaning what each state receives when AGI arrives depends only on who finishes first and how dangerous the arrival is, not on the record of conduct during the race. That assumption is realistic for now because no institution exists that conditions post-AGI outcomes on a verified record of pre-AGI behavior.[11] Instruments that condition post-AGI outcomes on conduct during the race (such as forfeitable stakes,

[11] Tools of economic statecraft, such as trade sanctions, could, in principle, continue after AGI arrives and punish pre-AGI violations to such an arrangement. However, these tools would need to be tied to pre-AGI behavior to deter defections from the arrangement but currently are not. Additionally, the sanctions must remain credible and consequential after AGI arrival.

benefit-sharing arrangements that shrink the gap between winning and losing, or prepositioned structures that a violator would lose) are excluded from the model by assumption and are the same instruments that would survive the limit. The "Prepare Instruments Early and in Self-Executing Form" subsection on page 16 describes what such instruments may look like.

Why Restraint May Still Fail

Our discussion so far concerns what oversight must deliver, which corresponds with Stage 2 of the game-theoretic model. In this section, we ask whether states will build the required oversight infrastructure, which is the decision made in Stage 1 of the game. Overall, the model suggests that even when the needed oversight is affordable and when states believe that it is worth building, each state's willingness to build such infrastructure depends on its forecast of the other's willingness, and mutual doubt can settle both on choosing an opaque world rather than a legible one.

The Oversight Game

The deterrence condition (Equation 3) assigns a required detection probability to each potential defection on the menu of defections, and the detection technology discussed in this section converts probability into a required oversight level, $\bar{x}$, the minimum infrastructure that deters every covered covert defection. Restraint requires both states to clear this threshold level of domestic oversight. If one state builds its oversight capacity and makes its AI ecosystem transparent while the other state does not and remains opaque, both states will revert to racing because the transparent state cannot rely on its rival's restraint—it cannot be checked. Therefore, the pair of Stage 1 choices—build the required capacity ($\bar{x}$) or remain at the domestic baseline—forms a two-by-two coordination game.[12]

Two Worlds

The Stage 1 game has two stable outcomes. We call the first the *legible world*. Here, both states build $\bar{x}$, every covered defection is deterred, and restraint holds for as long as the timing results discussed earlier allow. The other is the *opaque world*, where neither state builds the required oversight, some defections cannot be verified, and both states end up racing.[13] Both worlds are self-reinforcing because, against a rival that is building oversight capacity, building one's own oversight capacity is the best response, and against a rival that is not, saving the cost of building such capacity is the best response. Whether the legible world is even worth building depends on the comparison between the costs of building the required oversight capac-

This depends on the nonviolating state retaining the capacity and incentive to impose trade sanctions. If AGI confers a sufficiently decisive advantage, conventional sanctions may no longer impose a meaningful penalty. While the model does not include alternative policies, such as forfeitable stakes and prepositioned instruments (which may include but are not limited to trade sanctions), we do explicitly consider how such policies interact with the model because they directly tie AI development behavior to post-AGI outcomes.

[12] The baseline is not the absence of AI governance. Oversight in the model is measured relative to what states already choose to do for domestic reasons, so choosing zero means that the state has not *added* anything above what its domestic policies require. The coordination problem arises only when the increment that states need to add is actually required for the verification regime to hold. If unilateral collection and the oversight that states maintain for domestic purposes already deter the entire covered menu of defections, the required additional oversight is zero, and there is nothing on which states must coordinate.

[13] This simple two-world example assumes that at least one covered defection is not already self-deterred and not already covered by unilateral collection. If the entire menu of defections were deterred without any additional oversight, restraint would hold with no institutional investments, and no coordination problem would arise.

ity $c(\bar{x})$ and what that oversight capacity buys in terms of cooperative surplus S. When $c(\bar{x}) < S$, the legible world is strictly better for both states.

The Opaque World Can Prevail Even When Both States Would Prefer the Legible One

In the model, building oversight capacity only pays off if the rival builds the required level of capacity as well. Not building anything has the same payoff regardless of what the rival does. Under risk dominance (Proposition 4 in the appendix),[14] building the required capacity is the safer bet when

$$c(\bar{x}) < \frac{S}{2}. \tag{6}$$

The intuition is that a state that is uncertain of whether its rival will build capacity, weighing the two possibilities evenly, gains the surplus S only half of the time. So, the cost must be less than half the surplus to be worth the risk of building the required oversight. When $c(\bar{x})$ lies between $\frac{S}{2}$ and S, the legible world technically is better for both states, but the players' doubt about their rival building the required capacity means that they both choose not to build, and the opaque world becomes the equilibrium.

This is the formal skeleton of a familiar argument, which goes along the lines of *perhaps it would be best if no one pressed ahead, but since we cannot be sure the other side will hold back, we must press ahead ourselves.* This sentiment is found in essays written by Anthropic's CEO Dario Amodei (2026) and OpenAI's CEO Sam Altman (2024). The model helps identify the conditions under which the argument is correct. When restraint fails because the required level of oversight capacity is sufficiently high ($\bar{x} > 1$), the problem is structural because no oversight system could realistically be built that would make the binding defection detectable enough to deter. When restraint fails because $c(\bar{x}) > \frac{S}{2}$, the oversight infrastructure is worth building, but each side doubts that the other will build it, so the problem is one of coordination. In this case, anything that improves each side's forecast of the other's choice can help the states coordinate. We return to this idea in the upcoming "Use Reciprocal Demonstration to Select the Legible Equilibrium" subsection on page 17.

Unequal Costs of Legibility

The baseline model is symmetric, meaning that we assume the two states have similar beliefs and attributes. However, this assumption masks the cost of legibility to different political systems. Making an AI industry legible may require producing records that outsiders can check, enabling insiders to report failures, and tolerating external scrutiny of sensitive assets. States have different tolerances for this level of legibility.

To see how allowing for differing legibility costs influences the model, suppose the costs differ, so state i pays $c_i(\bar{x})$. The legible world exists only if the condition $c_i(\bar{x}) < \frac{S}{2}$ holds for both states, so the constraint binds for whichever state finds oversight more expensive, politically or administratively. The model then suggests that the feasibility of mutual restraint is determined by which rival is more opaque, and no amount of transparency by the more open state can substitute for that.

Two implications follow. First, a state that wants restraint has a self-interested reason to reduce its rival's cost of oversight, through technical assistance, shared monitoring platforms, or financed verification infrastructure. Within the model, this is an informal extension, since the model does not provide a way for states to make payments to one another, but the direction of the conclusion does not depend on adding them. Second, it suggests that the confidentiality-preserving verification technologies discussed in Baker et al.

[14] The concept of *risk dominance* is a game-theoretic method used to determine which equilibrium doubtful rivals choose when multiple equilibria are available. Risk dominance is a benchmark for what a player would do under strategic uncertainty, requiring that players choose the action that is safer against an unpredictable opponent.

(2025) matter most for the high-cost state because they shrink the gap between what verification must reveal and what a closed system can afford to show.[15]

Strategic Implications and Limitations

To show what the model suggests about verification architecture, start from the menu of defections judged possible, and for each candidate, ask the following five questions:

1. What immediate gain, positional gain, and added risk does the defection create?
2. What evidence would it leave in compute, power, procurement, facilities, personnel, cyber, or public data?
3. Who controls that evidence: the cheater, the victim, a private firm, a third country, or an intelligence service?
4. What sensitive information would be collected or shared and what does that evidence expose?
5. How much of the needed detection comes from unilateral collection and how much requires the defector's own oversight infrastructure?

These questions translate the model's conceptualization of verification into something more concrete for strategy. They also suggest that detection alone is not sufficient to deterring defections, because getting caught only matters if it costs the cheater something of value.

One caveat to the following discussion is that the model produces requirements for what an instrument must accomplish in order for restraint to survive a given failure mode but does not include specific policy instruments. The specific policies or strategies discussed in the subsequent sections are candidates that plausibly meet such requirements and are offered as notional illustrations rather than as findings.

Aim Verification at the Positional Defections

This strategy follows from Proposition 2 (in the appendix), which separates the defections that a state's own exposure to catastrophe already deters from those that require detection and external punishment. Risk-creating defections partly police themselves through the defector's own exposure to catastrophe. Purely positional defections, which change who arrives at AGI first but not the underlying risk of a catastrophic transition, are deterrable only by detection. The binding oversight requirement is therefore set by the high-position, low-risk rows of the menu of defections (notionally shown in row 3 of Table 1), and that is where verification design effort belongs. In practice, those rows may be harder to monitor reliably and could include such defections as weight exfiltration, secret algorithmic progress, and internal automated R&D, where the evidence is softest, and compute accounting helps least. The model's main demand for empirical work is to populate the menu of defections with credible estimates of each item's flow value, positional value, risk loading, and exposure to oversight because the binding item cannot be identified by intuition alone.

[15] The model prices the feasibility of restraint across political systems but does not price the desirability of particular winners and has no vocabulary for the judgment that AGI developed inside one political system is safer for the world than AGI developed inside another. A reader who holds that judgment is not contradicting the model. That reader is supplying parameters from outside the model, in the manner described in the "Two Empirical Assumptions" subsection on page 10, and the resulting verification requirements will sit at the demanding end of the range. What the model rules out is only the assumption that such judgments make verification unnecessary.

Of course, nothing requires states to form a single arrangement, and they may instead opt to build several smaller arrangements, each with its own set of covered defections. Separating issues by governing catastrophe-relevant conduct differently from positional conduct, such that a failure to detect defections in one agreement does not void the entire relationship, could also be considered. However, the model does not evaluate this case directly, and further research is required to determine the conditions under which breaking up restraint arrangements can actually deliver the mutual restraint that states desire.

Use Risk Assessments to Calibrate Verification Requirements

Shared risk assessment can help states determine how much of the incentive to defect from an arrangement is already offset by the expected danger of racing and how much must be offset by verification and punishment. A more optimistic assessment, suggesting that AGI poses little risk, raises the burden on verification because a state would expect to incur less harm from its own defection and to lose less when the states return to racing. A more pessimistic assessment lowers that burden. Formally, detection and perceived catastrophic loss reinforce each other, as shown in Proposition 2(iii) in the appendix.

Prepare Instruments Early and in Self-Executing Form

This strategy follows from the scope of Proposition 3 in the appendix. Near the arrival of AGI, the only quantities that still deter cheating are the ones that act on terminal payoffs. The requirement is therefore an instrument that conditions post-arrival outcomes on race conduct, built before it is needed. Instruments of economic statecraft, including export controls, tariffs, and market access withdrawal, are commonly used to punish state behavior and could apply here as well. Another candidate is a forfeitable restraint bond, which might involve cash, treasuries, or escrowed equity posted with a custodian and forfeited on verified violation, which would explicitly tie pre-AGI behavior to post-AGI terminal payoffs.

An obvious objection is that if reaching AGI first confers anything like a decisive advantage, that prize dwarfs any collateral a state or lab would plausibly post, and a winner powerful enough to claim that prize may be powerful enough to recover (by cyber means or otherwise) whatever it forfeited. The objection sets the design requirement for such an instrument. The stakes must be denominated in post-arrival standing rather than in cash, and the forfeiture must be self-executing rather than adjudicated after the fact. Plausible instruments include joint ventures whose governance and revenue rights a violator loses, physically distributed ownership of compute that no single party can repossess quickly, and legal structures prepositioned so that unwinding them is slow and visible.

Benefit-sharing arrangements that compress the spread between winning and losing work on the same terminal margin. These types of arrangements might redistribute what accrues to the player who finishes first through equity, access to resulting capabilities, access to AGI's economic output, or even joint ownership of the technology. In the model, the gap between finishing first and second is what drives positional defections, so compressing this margin lowers the detection an arrangement needs to survive. Such benefit-sharing arrangements might be less attractive to a state that sees itself as being sufficiently far ahead to believe that it is much more likely to finish first, which may make benefit-sharing most attractive for rivals who see themselves as being closely matched. Additionally, each must believe that the commitment would survive AGI arrival. Ultimately, an arrangement that depends on a winner's goodwill after that race is over may not survive post-AGI, so arrangements that are sufficiently costly for the eventual winner to unwind are likely the type required.

Treat Domestic AI Governance as Security Architecture

Compute accounting, licensing, facility registries, evaluation records, incident reporting, whistleblower protections, and assurance audits are usually debated as domestic risk-management tools and priced against domestic benefits. The model adds that they are also the infrastructure that makes a state's restraint claims believable, and their absence is what makes a rival's doubt rational. One concrete form such capacity could take is an assurance market for frontier AI, in which independent auditors examine whether labs keep the records and controls that make covert defection observable, issue public opinions, and secure sensitive details in confidential annexes, much as financial audits do. This approach is closely related to the proposal in Ball (2025) for government-authorized private bodies that certify AI developers against safety and security standards and conduct periodic audits. Paired with confidentiality-preserving verification technology (Baker et al., 2025), such a system lowers the security cost of transparency and converts compliance into a demonstrable public fact.

Use Reciprocal Demonstration to Select the Legible Equilibrium

If restraint fails in the coordination game (where oversight is affordable, but doubt means that the states select the opaque world), the remedy will be anything that improves each side's forecast of the other. The model cannot rank such devices but does specify that they must create verifiable evidence that the other side is willing and able to make itself checkable. One notional device is a reciprocal slowdown ladder among frontier AI labs, where a sequence of escalating, verified demonstrations of restraint, beginning with something as small as a synchronized one-day pause on covered training runs and escalating only after each rung is completed and verified by all participants. The ladder could produce a shared record of which actors can demonstrate restraint when asked, rung by rung, and each completed rung shifts every participant's forecast of its rival toward the legible equilibrium.

Limitations

This section collects and discusses the assumptions in the model, ordered by how much each could change the conclusions, with a judgment in each case about the direction of the bias.

States do not diverge in their beliefs. The model gives both states the same parameters and beliefs, which separates the enforcement problem from the inference problem. One objection to this assumption might be that rivals disagree about *everything* that matters in this context, especially about timelines to AGI (h) and severity of the transition (L). Additionally, beliefs within states are not constant, and political shifts or elections could cause the arrangement to fail (or succeed) for reasons outside this model. Divergent beliefs mean that the states disagree about how much oversight a cooperative restraint arrangement requires and about how much time the arrangement has left, and a state that suspects its rival of privately holding a low L should rationally discount that rival's restraint.

The common beliefs assumption primarily means that states agree on how much verification capacity is required. Adding different player types would make the Stage 1 oversight choice a signaling problem as well as a coordination problem. The amount of oversight that a state built would itself convey something about what the state believed regarding the danger of and the timeline to AGI, so the choice of domestic investment would be read as an informative signal as well as a capability. That extension is the natural next iteration of this analysis.

Cooperation is all or nothing. The baseline model treats an arrangement as something that either holds or collapses into permanent racing. However, real arrangements can be partial in the sense that they might collapse in some specific area and hold in others. Allowing for such partial arrangements would soften the

model's binary structure and, in particular, would make the exact coordination threshold conditional on how much the arrangement can be partitioned. As noted earlier, a partial-cooperative version of the model is a natural extension of this work.

Detection is perfect. The model assumes that there are no false positives, and every detection is publicly attributable. Real verification might produce ambiguous signals and contested attribution. The known consequence of adding noise (Green and Porter, 1984) is that punishment phases occur on the equilibrium path even when no one has cheated, restraint becomes costlier to hold, and finite punishments replace permanent ones. As a result, the conditions under which the cooperative restraint arrangement holds would become more demanding, and a new policy object (e.g., a dispute-resolution procedure) would become important.

States control their AI industries. In the model, a state's AI industry is compliant and does what the state decides. If labs have their own payoffs and private information, a state might end defections through actions that domestic labs took that the state failed to police. If domestic labs can go against the state's explicit choices, oversight serves both as a domestic control and as international transparency. Future work should incorporate the oversight choice in Stage 1 by allowing states to coordinate with domestic frontier AI labs.

Oversight is frozen after the first stage. Fixity is what makes oversight a commitment in the model. However, real oversight can decay or be hollowed out over time. If dismantling is observable, it is a signal that the model does not exploit. If dismantling can be hidden, oversight's commitment value falls toward zero. A dynamic version could exhibit cascades in which deteriorating international conduct erodes domestic governance and vice versa. The analysis suggests that interpreting growing nervousness inside a restraint regime as an early warning rather than as noise; however, confirming that as a dynamic prediction requires the dynamic model.

Risk and prizes ignore history. The probability of catastrophe and terminal payoffs depend on conduct in the period during which AGI arrives. If risk instead accumulates over the whole race, the self-deterrence effects could strengthen because past cheating raises the cheater's own permanent exposure. History-dependent terminal payoffs would be a deeper relaxation of the assumption.

No denial, no sabotage, no third parties. The set of choices that states can choose from in the model does not include a way to slow or disrupt a rival's AI development program, so the model is silent on deterrence-by-denial and on how restraint interacts with suppression postures. Third countries are similarly absent, though they may host the compute, talent, and corporate structures through which defections route their actions, and their inclusion would alter both detection technology and the meaning of domestic oversight. These are the model's connections to the wider strategy debate, and they remain open.

The states are politically symmetrical. The asymmetric-cost extension discussed in the "Unequal Costs of Legibility" subsection on page 14 captures differences in the price of legibility across states but does not capture divergent lab-state relations, different domestic audiences for compliance, or the judgment that AI development inside one political system is more dangerous than inside another.

Conclusion: A Strategic Test

Our central claim is that restraint feasibility is produced, in part, by what rivals do domestically to make their own AI ecosystems legible and verifiable. The results of the analysis suggest that (1) states' beliefs about the danger that AGI might pose and about how significant a strategic advantage arriving at AGI first might confer, (2) how near the states perceive AGI arrival to be and whether punishment for defecting from the arrangement holds meaningful weight, and (3) the costs that each state would bear, together with what each rival expects the other to do, determine whether the relevant domestic capacity is built in the first

place. These results suggest that strategists should consider the following four questions about any proposed restraint arrangement:

- First, what is on the menu of defections, and which items act as binding constraints on whether mutual restraint can be maintained? If the menu is dominated by risk-creating actions whose danger the defector would bear, restraint may need less verification than intuition suggests, and shared risk assessment could provide much of the required deterrence. If the menu is dominated by positional, hard-to-detect actions, restraint is likely more bound by verification limits and is more likely to fail even when both sides sincerely prefer it.
- Second, is the binding detection requirement reachable at acceptable cost? This is where the transparency-security trade-off and the technologies that soften it decide the answer and where the requirement lands hardest on the more opaque rival.
- Third, how much time is left before AGI arrives? If AGI arrival is perceived as close, detection-backed punishment could already be losing its purchase, and the instruments that can still deter cheating are the ones tied to post-arrival outcomes. Whether those exist at the moment they are needed depends entirely on whether they were built earlier.
- Fourth, if restraint is feasible, will states select it? Affordable oversight can still lose to mutual doubt, and in that situation, demonstration and sequencing are critical to designing effective arrangements.

The model presented in this report is not an argument for choosing restraint in AGI development. Instead, the discussion here is an account of what must be in place for restraint to remain a viable option for states that want it. That means identifying the relevant defections, making legibility less costly, building oversight capacity while it can still deter defections, and developing tools that reach beyond AGI arrival. A state that does these things has not bound itself to restraint but has preserved its ability to choose restraint if required.

APPENDIX

Technical Discussion of the Formal Model

Players, Timing, and Termination

There are two states, indexed as $i \in \{1,2\}$. The model is symmetric in the baseline, and the propositions note where asymmetry matters.

Play begins with each state simultaneously choosing an oversight level $x_i \in [0,1]$, which is publicly observed. *Oversight* is defined as regulatory infrastructure applied to the state's own AI industry. Its cost is $c(x_i)$, where c is continuous and strictly increasing, with $c(0) = 0$. The cost is incurred once, at the moment of choice, and is measured in the same present-value units as the continuation payoffs described in the "Payoffs" subsection on page 22, so the two are directly comparable. Let $x_i = 0$ represent the level of oversight a state would choose for purely domestic reasons, and c measures the net additional burden beyond that point. Oversight, once built, is fixed for the rest of the game.

After oversight levels are observed, the states play a repeated interaction game in periods $t = 1, 2, \ldots$ with a common discount factor of $\delta \in (0,1)$. The game has a random end. In each period, after actions are taken, AGI arrives with probability $h \in (0,1)$, ending the game and triggering the terminal payoffs for both players. The hazard h is constant and exogenous. The parameter h should be read as the perceived proximity of AGI arrival, and the "Proximity and the Enforcement Horizon" subsection on page 12 treats it as the central comparative-static variable.

The sequence of actions within a period is as follows:

1. Both states choose actions simultaneously.
2. Flow payoffs accrue.
3. With probability h, AGI arrives. Terminal payoffs are realized as a function of this period's action profile, and the game ends.
4. If the game did not end, any covert defection played this period is detected with probability given by the detection technology (as described in the "Detection and the Role of Domestic Oversight" subsection on page 23). Detection is public and attributable to the defector. Open racing is observed with certainty.
5. Play moves to the next period.

Actions

In each period, each state chooses an action from

$$\mathcal{A} = \{R\} \cup \{N\} \cup D,$$

where R is *restraint*: a state's compliance with the cooperative arrangement, whatever its institutional form. N is *open racing*: a state's overt abandonment of restraint, visible immediately to the other player. D is a finite *menu of covert defections*: ways of cheating while ostensibly complying with the cooperative arrangement.

Each covert defection $d \in D$ is characterized by three numbers and a function:

- $g_d \in \mathbb{R}$ is the flow gain per period of cheating, relative to compliance.
- $\pi_d \in [0, \frac{1}{2}]$ is the shift in probability of being first if AGI arrives while the cheating is in progress.
- $\rho_d \in [0, 1 - r_0]$ is the added probability that the transition is catastrophic if AGI arrives while the cheating is in progress (the upper bound keeps probabilities below one; r_0 is defined in the "Payoffs" subsection below).
- $q_d(\cdot)$ is the per-period detection probability (described in the "Detection and the Role of Domestic Oversight" subsection on page 23).

It is convenient to treat open racing as a member of the extended defection menu $\overline{D} = D \cup \{N\}$, with the characteristics (g_N, π_N, ρ_N) in the same ranges and detection probability identically set at one. Concealment is itself costly and limits scale, so the covert variant of an action will typically carry a lower g and lower π than its overt counterpart. The menu accommodates any such pattern. Nothing in the analysis requires any defection to be purely of one kind, and whether a given real-world action is mostly positional or mostly risk-creating is an empirical question that we return to later.

Payoffs

Flow payoffs. While the game continues, mutual restraint yields each state the payoff w per period. Mutual open racing yields each state w_N per period, with $\Delta w \equiv w - w_N > 0$. The surplus Δw collects the pre-arrival gains that restraint protects, such as unimpeded trade in inputs, shared safety research, lower security expenditure, and the absence of sabotage costs. A state playing a defection $d \in \overline{D}$ against a compliant rival receives $w + g_d$ in that period.

Terminal payoffs. When AGI arrives, one state finishes first. Under any symmetric action profile, each state is first with probability ½. If the arrival period's profile is (d, R), the defector is first with probability ½+ π_d. The first-mover prize is A and the other state receives B, with $A > B \geq 0$. The spread $A - B$ measures how decisive being first is, or the *decisive strategic advantage*.

Independently of which state is first, the transition may be catastrophic. Conditional on arrival, catastrophe occurs with probability r_0 under mutual restraint, r_N under mutual racing, and $r_0 + \rho_d$ under (d, R), where $\Delta r \equiv r_N - r_0 > 0$. Catastrophe costs each state $L > 0$. The parameter L is the priced severity of a failed transition, and a state optimistic about its coexistence with AGI holds a lower value of L. Risk depends on the survival period's conduct only. This is a simplification (risk, in fact, accumulates) and is flagged in Table A.1 (assumption A3) and in the "Limitations" section of the main body of this report.

Flow and terminal payoffs for profiles not specified above (the victim's flow payoff during an undetected defection, payoffs under double defection, and so on) may be anything, subject to the requirement that every flow and terminal payoff is a finite real number. None of the unspecified entries enter any equilibrium condition. Proposition 3 uses only finiteness.

It is useful to define, for any action profile $a = (a_1, a_2)$, the augmented per-period payoff as $\hat{u}_i(a) \equiv u_i(a) + h\,T_i(a)$, where u_i is the flow payoff and T_i is the expected terminal payoff conditional on AGI arrival under profile a. The augmented payoff bundles the per-period flow payoff with the hazard-weighted consequence of the game ending during the period. The three augmented payoffs that the analysis uses are as follows:

$$\hat{u}(R,R) = w + h\left[\tfrac{(A+B)}{2} - r_0 L\right],$$

$$\hat{u}(N,N) = w_N + h\left[\tfrac{A+B}{2} - r_N L\right], \text{ and}$$

$$\hat{u}_{dev}(d) = w + g_d + h\left[\tfrac{A+B}{2} + \pi_d(A-B) - (r_0 + \rho_d)L\right],$$

where $\hat{u}_{dev}(d)$ is the deviator's augmented payoff when it plays $d \in \overline{D}$ against a compliant rival.

Detection and the Role of Domestic Oversight

Formally, while state i plays covert defection d, the defection is detected at the end of the period (if the game has not ended) with probability $q_d(x_i)$, independently across periods, where $q_d:[0,1] \to [0,1]$ is continuous and strictly increasing. One example of this is

$$q_d(x) = q_d^0 + \kappa_d x,$$

where $q_d^0 \geq 0$ is the floor that espionage and national technical means provide, regardless of the rival's institutions, and $\kappa_d \geq 0$ measures how exposed defection d is to oversight infrastructure. Compute-intensive cheating has high κ because training runs leave procurement, energy, and facility footprints that registries and audits surface. Espionage-type cheating, such as exfiltrating a rival's model weights, has κ near zero because it is detected by the victim's security systems rather than by the thief's transparency. The propositions in this appendix use only monotonicity and continuity, not the linear form shown above. In "The Oversight Game" section on page 32, we note how state-specific technologies would change these results.

We assume that detection has no false positives. A detection event occurs only if a defection actually has occurred and if it identifies the defector. Real verification disputes are murkier, and the "Limitations" section on page 17 lists noisy attribution as a priority extension of the analysis described in this report. The argument of q_d is the defector's own oversight capacity. Choosing higher x_i is choosing to make one's own future cheating detectable. That is what gives oversight its strategic characteristic, and "The Oversight Game" section shows why a state might want it.

Assumptions

Table A.1 lists each assumption that disciplines the model. The table provides the reason for each assumption and what changes if the assumption fails.

TABLE A.1
Table of Assumptions

Number	Assumption	Why	If It Fails
A1	**The racing dilemma.** (i) Against an opponent playing N, no action gives a higher augmented payoff than N: $\hat{u}_i(N,N) \geq \hat{u}_i(a,N)\ \forall a \in \mathcal{A}$. (ii) Mutual restraint beats mutual racing: $\hat{u}(R,R) > \hat{u}(N,N)$. Part (ii) is implied by the sign conventions of the "Payoffs" subsection ($\Delta w > 0, \Delta r > 0, L > 0$) and is listed to make the dependence visible.	(i) makes permanent mutual racing a credible fallback: Once the rival races openly, matching the race is a best response, since covert actions are tuned to exploit a compliant rival and lose their edge against an open racer. (ii) is what makes restraint worth discussing.	If (i) fails, some action beats open racing even against an open racer, the punishment regime is misspecified, and the relevant fallback must be recharacterized before any of the analysis proceeds. If (ii) fails, there is no surplus to protect and the model is moot.
A2	**Detection technology.** Per-period detection probability $q_d(x_i)$ is increasing and continuous in the defector's oversight, no false positives, public attribution, contemporaneous only (an undetected defection can never surface later): $q_N \equiv 1$.	Encodes verification-as-cooperative technology, with espionage as a floor. Contemporaneous detection keeps the public record a sufficient statistic for continuation of play.	False positives put punishment on the equilibrium path and make restraint costlier to sustain (Green and Porter, 1984), so conclusions about when restraint fails are conservative. Delayed detection would add a forensic channel that strengthens deterrence. Rival-sided detection investment (espionage as a choice) would partially substitute for the defector's oversight on low κ defections, softening the coordination problem of Proposition 4.

Table A.1—Continued

Number	Assumption	Why	If It Fails
A3	**Catastrophe.** Conditional on AGI arrival, probability r_0, r_N, or $r_0 + \rho_d$ by arrival-period profile, costing each state the same L.	The simplest structure in which conduct affects risk and the defector internalizes part of the risk it creates.	If risk accumulates over the whole game rather than reflecting arrival-period conduct, defection history matters, and self-deterrence strengthens, because past cheating raises one's own permanent exposure. The taxonomy of Proposition 2 survives with different magnitudes. If the defector expects to bear less than the common L, through winner protection or mitigation, the self-deterred class shrinks, and the endgame result of Proposition 3(iii) weakens accordingly.
A4	**Prize structure, history-independent.** One winner (A), one follower (B), $A > B \geq 0$; symmetric profiles give even odds; defection d shifts odds by π_d. Terminal payoffs depend on the arrival-period profile only, never on the record of past play.	The minimal structure in which cheating can buy position. History independence reflects the absence, today, of any institution that could condition on post-AGI settlement on the verified record of conduct during the race.	Lead size–dependent or conduct-dependent magnitudes change. History dependence is deeper: Instruments that condition terminal payoffs on detected cheating (bonds, benefit-sharing forfeitures, post-arrival retaliation) would survive the $h \to 1$ limit and weaken Proposition 3(ii). The "Prepare Instruments Early and in Self-Executing Form" subsection (page 16) builds on this observation.
A5	**Oversight fixed after Stage 1.**	Registries, licensing regimes, and trained inspectors take time to build, and dismantling them is observable. Fixity is what makes oversight a commitment rather than cheap talk.	If oversight can be quietly dismantled, its commitment value falls toward zero and the Stage 1 analysis of Proposition 4 weakens accordingly. Observable dismantling is itself a signal, which the present model does not exploit.
A6	**Continuation selection.** After Stage 1, the states play the cooperative equilibrium of Proposition 1 when it exists, and permanent mutual racing otherwise.	The most favorable selection for cooperation. It stacks the deck toward restraint on purpose.	Any less favorable selection shrinks the cooperative region, so results about fragility and failure are conservative. Richer selections with partial cooperation below the threshold would soften the binary payoff structure of the oversight game: The two-equilibria coordination structure is robust, but the exact $\frac{S}{2}$ threshold of Proposition 4(iii) is conditional on the binary selection.
A7	**Common parameters, point beliefs.** Both states know all parameters, including each other's.	Separates the enforcement problem from the inference problem.	Belief divergence about h and L would hold that rivals with divergent beliefs are a dangerous configuration. We discuss this in the "Limitations" section (page 17).

Two assumptions, $\Delta w > 0$ and $\Delta r > 0$, are substantive claims about the world in which the states operate. Specifically, that restraint protects the real flow value and that a racing world is more accident-prone at AGI arrival than a restrained one. In the limiting benchmark $\Delta w \to 0$, enforcement far from arrival loses its force (the "Proximity and the Enforcement Horizon" section on page 30 shows why), and with $\Delta r = 0$, the complementarity in Proposition 2(ii) disappears. These two signs are the model's main empirical commitments.

The Continuation Game Is a Standard Repeated Game

Lemma 1. Fix oversight levels and consider the continuation game. For any strategy profile, each state's expected payoff equals the expected payoff of the same profile in an infinite-horizon repeated game with no termination, stage payoff $\hat{u}_i$, discount factor $\beta \equiv \delta(1-h) \in (0,1)$, and the same per-period public detection draw, so that histories and strategies in the two games coincide.

Proof. Let τ denote the arrival period. The arrival draw is i.i.d. across periods with constant hazard h, exogenous, independent of all action randomizations and detection draws, and realized after the period's actions. Hence, the game reaches period t with probability $(1-h)^{t-1}$, and conditional on reaching t, ends there with probability h. A state's realized payoff is

$$\sum_{t=1}^{\tau} \delta^{t-1} u_i(a_t) + \delta^{\tau-1} T_i(a_\tau) = \sum_{t=1}^{\infty} \delta^{t-1} \left[1\{\tau \geq t\}\, u_i(a_t) + 1\{\tau = 1\}\, T_i(a_t) \right].$$

Payoffs are bounded, and $\delta < 1$, so expectation passes through the infinite sum. Because the hazard coin flip is independent of everything the players see and do, and the period-t action is chosen before the period-t arrival draw, conditioning on $\{\tau \geq t\}$ or on $\{\tau = t\}$ carries no information about play, and the distribution of a_t given survival is the same as in the no-termination game. Therefore,

$$\mathbb{E}[\text{payoff}] = \sum_{t=1}^{\infty} \delta^{t-1} (1-h)^{t-1} \mathbb{E}\left[u_i(a_t) + h\, T_i(a_t) \right] = \sum_{t=1}^{\infty} \beta^{t-1} \mathbb{E}\left[\hat{u}_i(a_t) \right]. \square$$

Lemma 1 states that AGI imminence acts like impatience. A rise in the hazard h loads more weight onto the terminal consequence inside each period's $\hat{u}$ and shrinks the effective discount factor β, making the future matter less. Both effects are mechanical in the model, and we discuss the tension between these effects in the "Proximity and the Enforcement Horizon" section on page 30.

Because $\hat{u}$ is bounded and $\beta < 1$, changes to play far in the future alter payoffs by arbitrarily little (continuity at infinity). Two consequences follow. First, the equilibrium requirement must handle covert actions. A state that cheated without being detected knows something the rival does not, so the game has private histories, and the textbook notion of subgame perfection loses its grip because, after the first period, there are no proper subgames to test. One repair is the perfect public equilibrium (Abreu, Pearce, and Stacchetti, 1990), where both states use strategies that condition only on the public record, and no state can gain by deviating after any public history. The candidate profile below is public, and the proof of Proposition 1 (in the next section) verifies the stronger property that no state can gain by deviating after private histories either, so nothing rests on the distinction.

Second, with bounded payoffs and discounting, the one-shot deviation principle applies (Mailath and Samuelson, 2006). A public profile passes if and only if no state can gain by deviating in a single period at some history and conforming thereafter. The proofs below check this. Because the hazard is constant, the continuation game from any history is the same geometric problem, so the identity of Lemma 1 applies at every history.

Two regime values recur. If both states play R forever,

$$V^C = \left(\frac{\hat{u}(R,R)}{1-\beta} \right) = \frac{w + h\left[\frac{A+B}{2} - r_0 L \right]}{1-\beta}.$$

And if both play N forever,

$$V^N = \frac{\hat{u}(N,N)}{1-\beta} = \frac{w_N + h\left[\frac{A+B}{2} - r_N L\right]}{1-\beta}.$$

The cooperative surplus is their difference:

$$S \equiv V^C - V^N = \frac{\Delta w + hL\Delta r}{1-\beta} > 0$$

and is positive by the sign conventions of the "Payoffs" subsection on page 22 (this is A1(ii)). The composition of S matters later, because the two parts respond differently to proximity and to a state's optimism about the transition to AGI.

When Restraint Is Self-Enforcing

The candidate arrangement is that both states comply, but if either state is ever caught cheating or openly races, the cooperative relationship is over.

The grim-restraint profile σ^C. Each state plays R in every period in which no detection event has occurred, and neither state has played N. If a detection event occurs or either state plays N, both states play N in every subsequent period.

Permanent reversion to mutual racing is a credible punishment in the model because mutual racing forever is itself an equilibrium by assumption A1(i), and it is the harshest punishment among arrangements that revert to mutual racing. The conditions below are therefore the weakest under which any racing-reversion arrangement holds. Proposition 3(ii) shows that the model's central negative result does not depend at all on this choice of punishment.

Proposition 1 (Self-Enforcing Restraint). Given oversight levels (x_1, x_2), the grim-restraint profile is an equilibrium of the continuation game, in the sense that after every history, public or private, no state can gain by deviating (in particular, it is a perfect public equilibrium), if and only if, for each state i and every defection $d \in \overline{D}$,

$$g_d + h\pi_d(A-B) \le h\rho_d L + \beta q_d(x_i)S. \qquad (IC_d)$$

Proof. By Lemma 1, work in the equivalent repeated game with stage payoff $\hat{u}$ and discount β. The profile σ^C conditions only on the public record (detection events and open racing), so every history places the public state in either C (no breach on record) or N (a breach on record). A state's own undetected past defections are private but affect neither the opponent's future play, which is measurable in the public record alone, nor any future payoffs or transition probabilities. By assumption A2, detection is contemporaneous, so an undetected past defection can never surface later, and risk and prizes depend on the arrival-period profile only (by assumptions A3 and A4). The continuation problem that a state faces at any history depends only on the public regime, so it suffices to check one-shot deviations in regime N and in regime C, and checks over private histories automatically.

Regime N. The prescription is N forever for both, regardless of further history, so the public state stays N no matter what either state does. A one-period deviation to any $a \neq N$ changes the deviator's current augmented payoff by $\hat{u}_i(a,N) - \hat{u}_i(N,N) \le 0$ by assumption A1(i). The continuation value from the next period onward is V^N whether or not the deviation occurred, because the regime is absorbing. The deviation is weakly unprofitable, so σ^C is immune to one-shot deviations in regime N with no condition required.

Regime C. The prescription is *R*. Conforming yields the current payoff $\hat{u}(R,R)$ and, since no detection can occur when no defection is played (no false positives, per assumption A2), the regime remains *C*, for a total value of

$$V^C = \hat{u}(R,R) + \beta V^C.$$

Consider a one-period deviation to $d \in \overline{D}$ (the set $\overline{D}$ exhausts $\mathscr{A} \setminus \{R\}$), followed by conformity. The current augmented payoff is $\hat{u}_{dev}(d)$. At the end of the period, conditional on the game continuing (the survival probability is already inside β), the defection is detected with probability $q_d(x_i)$, sending the regime to N; otherwise the regime remains *C*, and the deviator returns to compliance. The deviation value is

$$V_d^{dev} = \hat{u}_{dev}(d) + \beta\left[q_d(x_i)\, V^N + \left(1 - q_d(x_i)\right) V^C\right].$$

Subtracting

$$V_d^{dev} - V^C = \underbrace{\hat{u}_{dev}(d) - \hat{u}(R,R)}_{\text{current gain}} - \underbrace{\beta\, q_d(x_i)(V^C - V^N)}_{\text{expected future loss}}$$

from the expressions in the "Payoffs" subsection, the w, $\frac{A+B}{2}$, and $r_0 L$ terms cancel in the current gain:

$$\hat{u}_{dev} - \hat{u}(R,R) = g_d + h\pi_d(A - B) - h\rho_d L,$$

so

$$V_d^{dev} - V^C = g_d + h\pi_d(A - B) - h\rho_d L - \beta\, q_d(x_i) S \equiv \Gamma_d. \tag{A.1}$$

The one-shot deviation to d is unprofitable when $\Gamma_d \le 0$, which is (IC_d). Deviations must be unprofitable for every $d \in \overline{D}$ and both states. By the one-shot deviation principle, these conditions are necessary and sufficient. □

Remark. A natural worry is that the most profitable way to cheat is to keep cheating until caught, which the one-shot check might miss. To check this, let V_d^D be the value of defecting with d every period until detection. It satisfies $V_d^D = \hat{u}_{dev}(d) + \beta\left[q_d V^N + \left(1 - q_d\right) V_d^D\right]$, so

$$V_d^D - V^C = \frac{\hat{u}_{dev}(d) + \beta\, q_d V^N - \left(1 - \beta(1 - q_d)\right) V^C}{1 - \beta(1 - q_d)}.$$

In the numerator, write $\left(1 - \beta(1 - q_d)\right) V^C = V^C - \beta V^C + \beta\, q_d V^C$ and substitute $V^C - \beta V^C = \hat{u}(R,R)$. The numerator becomes $\hat{u}_{dev}(d) - \hat{u}(R,R) - \beta\, q_d(V^C - V^N) = \Gamma_d$, so

$$V_d^D - V^C = \frac{\Gamma_d}{1 - \beta(1 - q_d(x_i))},$$

with a positive denominator. Permanent cheating beats compliance when one-shot cheating does. The threshold is the same.

Which Defections Can Be Deterred, and What Must Oversight Deliver?

For each $d \in D$, define the required detection probability as

$$\overline{q}_d \equiv \max\left\{0, \frac{g_d + h\pi_d(A - B) - h\rho_d L}{\beta S}\right\},$$

so that (IC_d) holds exactly when $q_d(x_i) \geq \overline{q}_d$. The expression contains no state index because both states share the same detection technology and parameters. Call d self-deterred if $\overline{q}_d = 0$ and enforcement dependent if $\overline{q}_d > 0$. The two classes partition any menu by the sign of the numerator: $g_d + h\pi_d(A - B) - h\rho_d L$.

Since q_d is increasing and continuous, define the required oversight $\overline{x}_d$ as the smallest x with $q_d(x) \geq \overline{q}_d$, when such an x exists in [0,1], and set $\overline{x}_d = \infty$ otherwise. Restraint requires every covert defection to be deterred, so the binding requirement is

$$\overline{x} \equiv \max_{d \in D} \overline{x}_d,$$

together with the overt condition $\Gamma_n \leq 0$ (with Γ_d as in Equation A.1), which involves no oversight because open racing is always seen. The overt condition, $g_N + h\pi_d(A - B) \leq h\rho_d L + \beta S$, is the classic repeated game constraint and would be present in any model of this kind. The covert conditions are the ones this model adds.

Restraint is feasible at proximity h if $\Gamma_N \leq 0$ and $\overline{x} \leq 1$. Feasibility here is derived from the defection menu, the detection technology, and the stakes.

Proposition 2 (Deterrence Taxonomy and the Verification Complementarity). Fix h, and consider any $d \in D$.

i. *Self-deterred defections.* If $h\rho_d L \geq g_d + h\pi_d(A - B)$, then $\overline{q}_d = 0$: Defection d is deterred at any oversight level, including $x_i = 0$ with no espionage floor, weakly at exact equality and strictly otherwise.

ii. *Pure positional defections are enforcement-dependent at every level.* If $\rho_d = 0$ and $g_d + h\pi_d(A - B) > 0$, then $\overline{q}_d > 0$ for every value of L, and if $q_d(x_i) = 0$, then (IC_d) fails for every value of L. Fear of catastrophe alone, however extreme, cannot deter a purely positional defection that cannot be detected.

iii. *Complementarity.* Define the deterrence margin, $M_d \equiv h\rho_d L + \beta q_d(x_i)S - g_d - h\pi_d(A - B)$, so that d is deterred when $M_d \geq 0$. Then

$$\frac{\partial M_d}{\partial L} = h\rho_d + \beta q_d(x_i)\frac{h\Delta r}{1 - \beta} \geq 0 \text{ and } \frac{\partial^2 M_d}{\partial q \partial L} = \frac{\beta h \Delta r}{1 - \beta} > 0.$$

Catastrophe pessimism (higher L) strengthens deterrence through two channels: directly in proportion to the defector's own risk loading, ρ_d, and through enforcement in proportion to the detection probability, because the punishment is reversion to a world that is more dangerous by Δr. The cross-partial derivative is strictly positive because detection capacity and catastrophe pessimism are complements.

iv. *Optimism raises the verification burden.* Whenever $\overline{q}_d > 0$, $\overline{q}_d$ is strictly decreasing in L. Consequently, $\overline{x}$ is nonincreasing in L, and falling L both raises the required oversight within the enforcement-dependent class and moves every self-deterred defection with $\rho_d > 0$ and $g_d + h\pi_d(A - B) > 0$ into the enforcement-dependent class once L falls far enough.

Proof.

i. the numerator of $\overline{q}_d$ is nonpositive by hypothesis, so $\overline{q}_d = 0$ and (IC_d) holds at $q_d = 0$, with equality only in the knife-edge cases.

ii. With $\rho_d = 0$, (IC_d) reads as $g_d + h\pi_d(A - B) \leq \beta q_d(x_i)S$. The left side is positive by hypothesis. If $q_d(x_i) = 0$, then the right side is zero, regardless of L, so the condition fails for every L. The required probability $\overline{q}_d = \frac{[g_d + h\pi_d(A - B)]^+}{\beta S}$ is strictly positive for every L because S is finite for every finite L.

iii. Substitute $S = \frac{\Delta w + hL\Delta r}{1-\beta}$ into M_d and differentiate: $\frac{\partial M_d}{\partial L} = h\rho_d + \beta q_d(x_i) \cdot \frac{h\Delta r}{(1-\beta)}$, which is nonnegative, and strictly positive if $\rho_d > 0$ or $q_d(x_i)\Delta r > 0$. Differentiating again in q, treating $\overline{q}_d(x_i)$ as the varying argument (which is how oversight moves it): $\frac{\partial^2 \mathrm{M_d}}{\partial q \partial L} = \frac{\beta h \Delta r}{1-\beta} > 0$ since $\Delta r > 0$.

iv. Where $\overline{q}_d > 0$, write $\overline{q}_d = \Theta_d^{net} \equiv g_d + h\pi_d(A-B) - h\rho_d L > 0$. Then

$$\frac{\partial \overline{q}_d}{\partial L} = \frac{-h\rho_d \cdot \beta S - \Theta_d^{net} \cdot \beta \frac{h\Delta r}{1-\beta}}{(\beta S)^2} < 0,$$

since both numerator terms are nonpositive, and the second is strictly negative ($\Theta_d^{net} > 0$, $\Delta r > 0$). Across its entire domain, $\overline{q}_d$ is continuous and nonincreasing in L: strictly decreasing where positive, and constant at zero thereafter. Hence, for $L_1 < L_2$, the set $\{x : q_d(x) \geq \overline{q}_d(L_2)\}$ contains $\{x : q_d(x) \geq q_d(L_1)\}$, so $\overline{x}_d(L_2) \leq \overline{x}_d(L_1)$ under the convention $\overline{x}_d = \infty$ when the set is empty, and the maximum $\overline{x}$ over the finite menu is nonincreasing in L. For the class change: a self-deterred defection with $\rho_d > 0$ and $g_d + h\pi_d(A-B) > 0$ satisfies $h\rho_d L \geq g_d + h\pi_d(A-B)$ at the prevailing L. Lowering L strictly below $L_d^* \equiv \frac{g_d + h\pi_d(A-B)}{h\rho_d} > 0$ reverses the inequality, making $\overline{q}_d > 0$. Defections with $\rho_d = 0$ or with $g_d + h\pi_d(A-B) \leq 0$ never change class as L varies. □

Proposition 2 shows three things. First, the two classes partition any menu by the sign of a derived quantity, and the location of the boundary is a model output. The model takes no position on where real defections fall. However, assumptions do determine the direction in which these characteristics operate in the model. Specifically, under assumption A3, ρ_d can deter by making the defector bear its own added risk, and assumption A4 puts π_d on the temptation side. The threshold itself is derived, as well as its proximity scaling (the self-deterrence wedge grows with h, while g_d does not, which is what Proposition 3(iii) exploits), and the complementarity in Proposition 2(iii). Locating real defections on the menu is a measurement task, and Proposition 2 tells you what to measure: specifically, each candidate defection's flow gain, positional value, risk loading, and exposure to oversight. Table A.2 sketches how that exercise might turn out.

TABLE A.2

Example Defections and Their Relationships to the Model

Candidate Defection	g_d	π_d	ρ_d	κ_d (Oversight Exposure)
Covert training runs above declared compute thresholds	Moderate	High	Low to moderate	High (compute accounting, energy, and procurement audits)
Undeclared frontier AI lab or program	Moderate	High	Moderate	High (licensing, registries)
Exfiltration of rival's model weights	Low	Moderate	Low to moderate	Near zero (detected by victim's security, so deterrence rests on the espionage floor)
Safety evaluation shortcuts before frontier AI development	Moderate (speed)	Low	High	Moderate
Open breakout (renounce and race)	High	High	High	Irrelevant ($q_N = 1$)

The pattern in Table A.2 suggests that the binding constraint $\bar{x}$ is set by the high-π, low-ρ rows, and verification infrastructure should be designed against the positional defections, because the risk-raising ones partially police themselves.

Second, part (iii) adds an intuitive complementarity. Part (ii) states that fear of catastrophe cannot substitute for detection on positional defections. It can still bear on them, because higher L makes the fallback world worse, raises S, and so raises the price of getting caught. The channel is multiplied by $q_d(x_i)$, so it pays out only where detection capacity exists. Shared risk assessment and verification machinery are complements, so each raises the return on the other.

Third, part (iv) is the result if a state becomes convinced that the AGI transition is survivable (low L); it is harder to hold the state in restraint on every margin. Its self-deterrence shrinks, the punishment it fears shrinks, and the oversight level needed to compensate rises. Estimates of transition severity are therefore load-bearing inputs to verification design, and deserve the scrutiny usually reserved for capability estimates.

Proximity and the Enforcement Horizon

The statements below are comparisons across stationary environments that differ in h. Reading them as a within-game trajectory of rising perceived proximity is heuristic, and we have suggested a nonstationary extension as a future research project.

Write the upper envelope of the enforcement wedge (detection probability is one, attained only by open racing) as follows:

$$E_{\max}(h) \equiv \beta S = \frac{(\delta(1-h)[\Delta w + hL\Delta r])}{1-\delta(1-h)}.$$

Proposition 3 (The Enforcement Horizon).

i. *The enforcement wedge vanishes:* $\lim_{h\to 1} E_{\max}(h) = 0$. As AGI arrival becomes imminent, the enforcement wedge vanishes even at certain detection, while the temptation and self-deterrence terms converge to $g_d + \pi_d(A-B)$ and $\rho_d L$, respectively.

ii. *Robustness to arbitrary punishment.* Fix any defection $d \in \overline{D}$ with $g_d + \pi_d(A-B) > \rho_d L$. There exists $\overline{h}_d < 1$ such that, for all $h > \overline{h}_d$, no Nash equilibrium of the continuation game (and hence no equilibrium under any refinement) involves both states playing R on the equilibrium path in every period. This holds for every oversight pair, every detection technology, and every punishment delivered through pre-arrival continuation play, grim or otherwise.

iii. What survives. Fix any $d \in D$ with $\rho_d L > g_d + \pi_d(A-B)$. There exists $\hat{h}_d < 1$ such that, for all $h > \hat{h}_d$, (IC_d) holds even at $q_d(x_i) = 0$. Because the menu is finite, the combined threshold $h^* \equiv \max\{\hat{h}_d$ over the class of (iii), $\overline{h}_d$ over the class of (ii)$\}$ is below one, and for all $h > h^*$, the classification is complete on generic menus (those with $\rho_d L \neq g_d + \pi_d(A-B)$ for every d. Defections with $\rho_d L > g_d + \pi_d(A-B)$ are deterred without any verification, and defections with the reverse strict inequality are undeterrable.

Proof.

i. As $h \to 1$, the numerator of $E_{\max}$ tends to $\delta \cdot 0 \cdot [\Delta w + L\Delta r] = 0$, and the denominator tends to one, so $E_{\max}(h) \to 0$. The temptation term $g_d + h\pi_d(A-B) \to g_d \pi(A-B)$, and the self-deterrence wedge $h\rho_d L \to \rho_d L$ by continuity.

ii. All stage payoffs are finite real numbers, so the augmented stage payoff is uniformly bounded: For all profiles a and all $h \in [0,1]$, $|\hat{u}_i(a)| \leq \max_a |u_i(a)| + \max_a |T_i(a)| \equiv K < \infty$, with K independent of h because $h \in [0,1]$ multiplies a bounded term. Hence, every feasible continuation

value, in any equilibrium, lies in $\left[-\frac{K}{1-\beta}, \frac{K}{1-\beta}\right]$, and the largest possible difference between any two continuation values is $\frac{2K}{1-\beta}$.

iii. Suppose a Nash equilibrium has both states playing R on an equilibrium path in every period. With no defections and no false positives, the on-path public record is deterministic, so on-path histories are reached with positive probability and play at them is constrained by the equilibrium condition. Consider state i deviating to d for one period at an on-path history and playing optimally thereafter. The deviation changes the current augmented payoff by $g_d + h\pi_d(A - B) - h\rho_d L$ and changes the discounted continuation value by, at most,

$$\beta \cdot \frac{2K}{1-\beta} = \frac{2K\delta(1-h)}{1-\delta(1-h)}$$

in absolute value, whatever the opponent's strategy prescribes after any history and any detection outcome. Equilibrium therefore requires

$$g_d + h\pi_d(A - B) - h\rho_d L \leq \frac{2K\delta(1-h)}{1-\delta(1-h)}. \tag{A.2}$$

As $h \to 1$, the left side tends to $g_d + \pi_d(A - B) - \rho_d L > 0$ by hypothesis, while the right side tends to 0. By continuity of both sides in h, there exists an $\overline{h}_d < 1$ such that Equation A.2 fails for all $h > \overline{h}_d$. For such h, the supposed equilibrium admits a profitable deviation, a contradiction. The argument never uses the credibility of off-path play, which is why it covers Nash equilibrium and everything stronger.

iv. The deterrence margin at zero detection is $M_d^0(h) = h\rho_d L - g_d - h\pi_d(A - B) = h\big(\rho_d L - \pi_d(A - B)\big) - g_d$. As $h \to 1$, $M_d^0 \to \rho_d L - \pi_d(A - B) - g_d > 0$ by hypothesis. By continuity, there exists $\widehat{h}_d < 1$, with $M_d^0(h) > 0$ for all $h > \widehat{h}_d$. For the combined threshold, the menu is finite, each $\widehat{h}_d$ in the class of Proposition 3(iii) and each $\overline{h}_d$ in the class of Proposition 3(ii) is undeterrable (its own $\overline{h}_d$ is exceeded), and Equation A.2 failing at the bound $\frac{2K}{1-\beta} \geq$ any feasible continuation loss implies that (IC_d) fails at every $q \leq 1$. On a generic menu, these two classes exhaust D. □

A knife-edge defection, with $\rho_d L$ exactly equal to $g_d + \pi_d(A - B)$, interpolates between the classes. Its margin at zero defection is $-g_d(1 - h)$, so with $g_d \leq 0$, it stays deterred at every $h < 1$ with no verification. With $g_d > 0$, both its net temptation and the maximal enforcement wedge vanish linearly as $h \to 1$, and deterrence at certain detection survives the limit exactly when $g_d \leq \delta(\Delta w + L\Delta r)$.

Remark (scope of the quantifier). The quantifier in Proposition 3(ii) is over punishments delivered through pre-arrival continuation play. Boundedness of the stage payoff is what the argument uses. Any scheme whose sting arrives before AGI does is discounted by $\beta \to 0$. Instruments that condition terminal payoffs on detected cheating, such as forfeitable bonds, benefit-sharing, or post-arrival retaliation promises, sit outside the action space per assumption A4. They are excluded by construction. The "Prepare Instruments Early and in Self-Executing Form" subsection on page 16 builds on this observation.

Remark (enforcement need not decay monotonically). Proposition 3(i) is a limit statement, and the path to the limit can be nonmonotone. Differentiating gives $E'_{\max}(0) = \frac{\delta[L\Delta r(1-\delta) - \Delta w]}{(1-\delta)^2}$, which is positive when $L\Delta r(1 - \delta) > \Delta w$. When the punishment's sting comes mostly from the riskiness of the fallback world rather than from lost flow value, moderate proximity strengthens enforcement before extreme proximity destroys it.

Remark (the far-from-arrival benchmark). At $h \to 0$, the self-deterrence wedge and the positional term vanish, so the (IC_d) collapses to

$$g_d \leq \delta q_d(x_i)\frac{\Delta w}{1-\delta}.$$

Deterrence far from AGI arrival rests on enforcement and only flow magnitudes matter. Restraint at long horizons is held together by the value of the ongoing relationship, in particular Δw. If restraint protects no flow value, distant horizon enforcement has nothing to work with.

The Oversight Game

Stage 2 told us what oversight must deliver, while Stage 1 asks whether states will build it. Under assumption A6, Stage 1 payoffs are as follows:

$$U_i(x_i, x_j) = \begin{cases} V^C - c(x_i) & \text{if } \min\{x_1, x_2\} \geq \bar{x} \text{ and } \Gamma_N \leq 0 \\ V^N - c(x_i) & \text{otherwise.} \end{cases}$$

The threshold structure is inherited from Proposition 1, where restraint requires each state's covert defections to be deterred, so it requires both oversight levels to clear $\bar{x}$. One transparent state and one opaque state results in fallback racing, because the transparent state cannot trust the opaque one, whatever its own virtues. Oversight is therefore a weakest-link good. If the two states face different detection technologies, the threshold would be state-specific, and every appearance of $\min\{x_1, x_2\} \geq \bar{x}$ in Proposition 4 becomes $x_i \geq \bar{x}_i$ for each i, with nothing else changing in Propositions 1 and 2.

Throughout this section, assume restraint is feasible and nontrivially demanding: $\Gamma_N \leq 0$ and $0 < \bar{x} \leq 1$. (If $\bar{x} = 0$, the espionage floor and self-deterrence wedge already deter the whole menu, restraint holds with no institutional investment, and there is no coordination problem. The fragility Proposition 4(iii) is specifically a product of an enforcement-dependent menu.)

Proposition 4 (The Oversight Game Is a Stag Hunt). Assume $c(\bar{x}) < S$. (Positivity of $c(\bar{x})$ is automatic, since $\bar{x} > 0$, and c is strictly increasing from $c(0) = 0$.)

i. Every oversight level outside $\{0, \bar{x}\}$ is strictly dominated: $x_i \in (0, \bar{x})$ is dominated by 0, and $x_i > \bar{x}$ is dominated by $\bar{x}$.

ii. In the game restricted to the undominated levels $\{0, \bar{x}\}$, there are exactly two pure-strategy Nash equilibria, both strict (each state's own oversight level is the unique best response to the other's choice): the legible world $(\bar{x}, \bar{x})$, where restraint is sustained, and the opaque world (0,0), where it is not. The legible world gives both states strictly higher payoffs. This is a stag hunt with two equilibria, one better for both, each safe only if the other side matches it.

iii. The legible world is risk-dominant if and only if

$$c(\bar{x}) < \frac{S}{2},$$

with neither equilibrium risk-dominant at exact equality. When $c(\bar{x}) \in \left(\frac{S}{2}, S\right)$, the opaque world is risk-dominant even though the legible world is better for both states.

Proof.

i. Fix any x_j.
   a. For $x_i \in (0, \bar{x})$: State i fails the threshold either way, so the regime indicator, and hence the gross payoff, is the same at $x_i = 0$ for every x_j, while the cost is strictly higher because c is strictly increasing. So, 0 strictly dominates.
   b. For $x_i > \bar{x}$: State i clears the threshold either way, so for every x_j, the regime outcome is identical to that at $x_i = \bar{x}$, while the cost is strictly higher. So $\bar{x}$ strictly dominates.

ii. The restricted game has a payoff matrix (the player's payoff appears first in each row of the table below):

	$x_j = \bar{x}$	$x_j = 0$
$x_i = \bar{x}$	$V^C - c(\bar{x})$, $V^C - c(\bar{x})$	$V^N - c(\bar{x})$, V^N
$x_i = 0$	V^N, $V^N - c(\bar{x})$	V^N, V^N

- Check $(\bar{x},\bar{x})$: Deviating to 0 changes the deviator's payoff by $V^N - (V^C - c(\bar{x})) = c(\bar{x}) - S < 0$ by hypothesis. Strict equilibrium.
- Check (0,0): deviating to $\bar{x}$ changes the deviators payoff by $(V^N - c(\bar{x})) - V^N = -c(\bar{x}) < 0$. Strict equilibrium.
- Check the asymmetric profiles: At $(\bar{x}, 0)$, the first state gains $c(\bar{x}) > 0$ by moving to 0, so it is not an equilibrium, and is symmetrically for $(0, \bar{x})$.

iii. For any symmetric two-by-two coordination game, the risk-dominant equilibrium is the one with the larger product of unilateral deviation losses (Harsanyi and Selten, 1988). The deviation loss at the legible world is $S - c(\bar{x})$ for each state; at the opaque world, it is $c(\bar{x})$ for each. The legible world is risk-dominant when

$$(S - c(\bar{x}))^2 > (c(\bar{x}))^2 \iff S - c(\bar{x}) > c(\bar{x}) \iff c(\bar{x}) < \frac{S}{2},$$

using $S - c(\bar{x}) > 0$, with equality of the products exactly at $c(\bar{x}) = \frac{S}{2}$. □

Risk dominance selects the equilibrium a player would choose under maximal strategic doubt, when it assigns even odds to the rival's two candidate actions. Concretely, against a rival equally likely to build or not its oversight capacity, building pays $\frac{1}{2}(V^C - c(\bar{x})) + \frac{1}{2}(V^N - c(\bar{x}))$ and not building pays V^N, so building is the better reply when $c(\bar{x}) < \frac{S}{2}$. That is the same condition the product criterion delivers in Proposition 4(iii). Risk dominance is the standard formal answer to which equilibrium is focal when trust is thin and has an independent foundation (Carlsson and van Damme, 1993). Between rival states with thin trust, risk dominance is a conservative prediction, since assumption A6 already resolves every other selection question in cooperation's favor.

An assistance corollary. Let costs differ such that state i pays $c_i(\bar{x})$. The legible world exists as a strict equilibrium if and only if $c_i(\bar{x}) < S$ for both states, so the constraint binds at whichever state finds oversight more expensive, politically or administratively. Therefore, a state that wants restraint has a direct interest in lowering its rival's oversight costs via technical assistance, shared monitoring platforms, financed verification infrastructure, and the like. Financing the rival is self-interested whenever the funder state's surplus covers its own cost plus the transfer, $S > c_i(x_i) + t$. This reading is an informal extension, because transfers are not in the action set, but its direction is forced by the existence condition.

References

Abraham, Lisa, Joshua Kavner, and Alvin Moon, *A Prisoner's Dilemma in the Race to Artificial General Intelligence*, RAND Corporation, RR-A4245-1, 2025. As of September 3, 2026: https://www.rand.org/pubs/research_reports/RRA4245-1.html

Abreu, Dilip, David Pearce, and Ennio Stacchetti, "Toward a Theory of Discounted Repeated Games with Imperfect Monitoring," *Econometrica*, Vol. 58, No. 5, September 1990.

Altman, Sam, "Who Will Control the Future of AI?" *Washington Post*, July 25, 2024.

Amodei, Dario, "The Adolescence of Technology: Confronting and Overcoming the Risks of Powerful AI," January 2026.

Baker, Mauricio, "Nuclear Arms Control Verification and Lessons for AI Treaties," arXiv, arXiv:2304.04123, April 8, 2023.

Baker, Mauricio, Gabriel Kulp, Oliver Marks, Miles Brundage, and Lennart Heim, "Verifying International Agreements on AI: Six Layers of Verification for Rules on Large-Scale AI Development and Deployment," arXiv, arXiv:2507.15916, July 21, 2025.

Ball, Dean W., "A Framework for Private Governance of Frontier Artificial Intelligence," arXiv, arXiv:2504.11501, April 15, 2025.

Brundage, Miles, Noemi Dreksler, Aidan Homewood, Sean McGregor, Patricia Paskov, Conrad Stosz, Girish Sastry, A. Feder Cooper, George Balston, Steven Adler, et al., "Frontier AI Auditing: Toward Rigorous Third-Party Assessment of Safety and Security Practices at Leading AI Companies," arXiv, arXiv:2601.11699v4, February 7, 2026.

Bureau of Industry and Security, U.S. Department of Commerce, "Establishment of Reporting Requirements for the Development of Advanced Artificial Intelligence Models and Computing Clusters," proposed rule, *Federal Register*, Vol. 89, No. 176, September 11, 2024.

Carlsson, Hans, and Eric van Damme, "Global Games and Equilibrium Selection," *Econometrica*, Vol. 61, No. 5, September 1993.

Coe, Andrew J., and Jane Vaynman, "Why Arms Control Is So Rare," *American Political Science Review*, Vol. 114, No. 2, 2020.

Fudenberg, Drew, and Andrew Koh, "Racing to Ruin," arXiv, arXiv:2607.27638, July 30, 2026.

Green, Edward J., and Robert H. Porter, "Noncooperative Collusion Under Imperfect Price Information," *Econometrica*, Vol. 52, No. 1, January 1984.

Harsanyi, John C., and Reinhard Selten, *A General Theory of Equilibrium Selection in Games*, MIT Press, 1988.

Heijnen, Pim, and Lammertjan Dam, "Catastrophe and Cooperation," *Dynamic Games and Applications*, Vol. 9, March 2019.

Ho, Anson, Tamay Besiroglu, Ege Erdil, David Owen, Robi Rahman, Zifan Carl Guo, David Atkinson, Neil Thompson, and Jaime Sevilla, "Algorithmic Progress in Language Models," arXiv, arXiv:2403.05812, March 9, 2024.

Larsen, Thomas, Romeo Dean, Brendan Halstead, Eli Lifland, Ryan Greenblatt, and Daniel Kokotajlo, "AI 2040: Plan A," AI Futures Project, July 2026.

Mailath, George J., and Larry Samuelson, *Repeated Games and Reputations: Long-Run Relationships*, Oxford University Press, 2006.

Pacing the Frontier, homepage, July 2026. As of July 11, 2026: https://www.pacingthefrontier.com/

Predd, Joel B., Benjamin Boudreaux, Matt Chessen, Beba Cibralic, Edward Geist, Kamaria Horton, Amanda Kerrigan, William Marcellino, Shanshan Mei, Jared Mondschein, Alvin Moon, and Tobias Sytsma, *A U.S. Strategy to Secure Geopolitical Advantage on an Uncertain Path to Superintelligence: Maintaining Freedom of Action*, RAND Corporation, PE-A5105-1, 2026. As of September 21, 2026: https://www.rand.org/pubs/perspectives/PEA5105-1.html

Root, Philip, *Equilibrium Strategies on the Path to Artificial General Intelligence*, RAND Corporation, PE-A4788-1, June 2026. As of September 3, 2026: https://www.rand.org/pubs/perspectives/PEA4788-1.html

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

U.S. House of Representatives, AI Incident Reporting Act, H.R. 9477, June 25, 2026.

Vaynman, Jane, and Tristan A. Volpe, "Dual Use Deception: How Technology Shapes Cooperation in International Relations," *International Organization*, Vol. 77, No. 3, Summer 2023.

Vaynman, Jane, and Tristan A. Volpe, "Competition and Collusion: How the AI Arms Race Can Motivate Governance," in Jim Mitre, Michael C. Horowitz, Natalia Henry, Emma Borden, and Joel B. Predd, eds., *The Artificial General Intelligence Race and International Security*, RAND Corporation, PE-A4155-1, September 2025. As of September 3, 2026: https://www.rand.org/pubs/perspectives/PEA4155-1.html

Zaidi, Waqar, and Allan Dafoe, "International Control of Powerful Technology: Lessons from the Baruch Plan for Nuclear Weapons," working paper 2021:9, Centre for the Governance of AI, Future of Humanity Institute, University of Oxford, March 2021.

About the Authors

Tobias Sytsma is an economist at RAND. He conducts research on topics around technological change, international trade, and economic security. He holds a Ph.D. in economics.

Alvin Moon is a mathematician at RAND. His research focuses on modeling and mathematical analysis across several domains, including AI, cryptography, supply chains, and workforce development. He has a Ph.D. in mathematics.