Strategic Transparency

L.C.R. Patell

Abstract

The United States and China may face an accelerating race towards developing advanced AI capabilities. In that context, this piece argues for*strategic transparency*: targeted information sharing between states that are engaging in a race to build superintelligence. A corollary of strategic transparency is that *securitization is not secrecy*. Even in regimes where both states pursue high secure AI development at the frontier, they can strategically release information to shift their rivals beliefs. Building on previous formal analysis of AI races, we endogenize signaling to frame the race as an information design problem. Drawing on that literature, we construct three increasingly complex models that demonstrate that strategic transparency can stabilize arms race dynamics and that states have access to a range of options for effective signaling.

Introduction

The race towards superintelligence resembles a textbook security dilemma. Both the United States and China fear that delay risks conceding a decisive advantage; each therefore accelerates.

Yet **securitization is not secrecy**: even in regimes where both states pursue highly secure national projects, they can strategically release information to shift their rival's beliefs. These shifts can contribute to strategic stability by dampening the pressure to accelerate domestic AI development.

That possibility generates a biting question:

> *How should the United States design evaluations of its frontier AI systems so as to minimize China's incentive to accelerate?*

If the literature and public discourse on a race towards advanced AI, however, there is significant disagreement concerning how much information to share – if indeed states should share *any* information.

Armstrong et al. (2016) presents a canonical model of an AI race in which teams compete in a winner-take all competition to develop the first "proper AI", which connotes superintelligence in the tradition of Bostrom (2014). Armstrong et al. do not model information sharing endogenously. Instead, they consider three different settings: *no information*, in which "every team is ignorant of their own or any other team's capabilities", *private information*, in which "each team knows its own capabilities, but not those of the other teams", and *public information*, in which "every team knows the capabilities of every other team". They do not consider which information sharing setting is best *from the point of view of the actors*; rather, they evaluate the settings from the point of view of safety from existential risk. Their models yield the conclusion that no information is the safest setting and that the relative safety of the private information and public information cases depends on the degree of *enmity*: the extent to which an actor loses utility when another team builds superintelligence first.

Stafford et al. (2022) likewise treats information sharing as exogenous. AI capabilities and "knowledge levels" – roughly, the technical knowledge an actor has that increases the probability with which they successfully implement superintelligence – are public information. The game is one of complete information: each actor knows the other's capabilities and knowledge level exactly. As a result, the question of strategic transparency does not arise.

Emery-Xu et al. (2024) generalizes the model in Armstrong et al. (2016) to investigate the role of information and uncertainty concerning actors' progress in AI research and capabilities. In this model, two states are competing to build a "significant military technology". Nature endows each state with a research capability level x_i, which is determined exogenously, and then states select an investment into safety s_i. After selection, a state either wins, or there is an AI-enabled disaster. Like Armstrong et al., the primary focus of this model is determining "disaster risk" – the "expected probability of disaster". They solve for the unique symmetric Bayesian Nash Equilibrium in the no information, private information, and public information states. Like Armstrong et al. (2016), Emery-Xu et al. (2024) finds that "the no information scenario is always safer than the private information scenario, while the relative safety of the public and private information scenarios depends on m", which is the *decisiveness parameter* in the logistic contest success function. While this model incorporates uncertainty concerning progress and capability levels, there is no endogenous information sharing. Again, the question of strategic transparency does not arise.

The conclusions of Armstrong et al. (2016) and Emery-Xu et al. (2024) might be taken to imply that states should not share information under certain conditions – namely, whenever the private information scenarios are safer than the public information scenarios. Nevertheless, crucial lacunae remain: information sharing remains an exogenous feature of the models, which do not countenance the possibility of signaling that is finer grained than full revelation.

To investigate the question of information sharing, this paper casts the problem as one of *information design.* Sections 1 and 2 solve a binary model in which a dangerous capability level is either reached $\theta = 1$ or not $\theta = 0$. Section 3 extends the model to a continuous capability space. Section 4 studies a *voluntary disclosure* game with selectively disclosed, verifiable evidence. These models suggest two high-level findings:

1. **Strategic transparency exists**. Designing model evaluations and sharing their results can contribute to stabilizing arms race dynamics.

2. **Options are available.** When states can acquire hard evidence, then voluntarily disclosing a subset can approximate the information design optimum.

1 A Binary Strategic Setting

This section defines the core model on which I will build throughout the rest of the paper.

Two strategic actors – the United States (**U**) and China (**C**) – choose whether to accelerate (A) domestic AI development or to pause (P) for the current decision period. The game is one-shot.

Let *capability* $\theta = \{0, 1\}$. If $\theta = 1$, a frontier capability has been reached (such as a model that can autonomously generate novel zero-days exploits at human expert level); $\theta = 0$ means that it has not.

States share a common prior that the US has reached a new capability level $p = P(\theta = 1) = 0.7$. Given the race dynamics at play, it seems reasonable to believe that states will think it more likely than not that capabilities will continue to progress.

Alignment difficulty $a \in \{L, H\}$. We can think of alignment difficulty roughly as a proxy for loss of control: ($a = H$) renders loss of control risk high, whereas ($a = L$) means that loss of control risk is low, but nonzero.

To simplify the initial modeling, I assume that $a = H$ and that states' common prior is $q = P(a = H) = 1$.

After observing (θ, a), **U** may run an evaluation procedure (eval) that produces a public signal $s \in \{0, 1\}$. Formally, an eval is an experiment, determined by a pair of error rates:

$$\alpha = P[s = 1|\theta = 1],$$

$$\beta = P[s = 1|\theta = 0].$$

States gain utility from the actions that **C** takes. **U**'s payoff is:

$$U_U(a_c) = \begin{cases} 0 & \text{if } \mathbf{C} \text{ accelerates,} \\ 1 & \text{if } \mathbf{C} \text{ pauses.} \end{cases}$$

When China accelerates, the United States incurs both the risk of falling behind and suffers from increased loss of control risk, which presumably constitutes a global externality. As a result, it is always better for the United States if China pauses.

C's payoff is:

$$U_C(a_c, \theta) = \begin{cases} 1 & \text{if } \mathbf{C} \text{ accelerates and } \theta = 1, \\ 0 & \text{if } \mathbf{C} \text{ accelerates and } \theta = 0, \\ 1 & \text{if } \mathbf{C} \text{ pauses and } \theta = 0, \\ 0 & \text{if } \mathbf{C} \text{ pauses and } \theta = 1. \end{cases}$$

The reasoning here stems from the twin risks that, *ex hypothesi*, motivate **C**: being dominated by the United States if it falls behind, and loss of control. We assume that loss of control – including the probability of extinction – is equally bad for **C** as being dominated, given our assumption that $a = H$.

Nature determines the value of θ and neither state initially observes the resulting state. **U** then selects an experiment and commits to publishing the realized signal s using an *honest* mechanism:

> **U** tells the truth, the whole truth, and nothing but the truth. In other words, he has committed to fully disclose all his private information".[1]

[1] Kamenica and Gentzkow (2011), p. 9.

C observes s, updates its beliefs according to Bayes' law, and chooses A or P. States then receive their respective payoffs.

C's posterior is given by:

$$\mu = P(\theta = 1|s).$$

In the base counterfactual, there is no informative signal, and so **C**'s posterior equals its prior:

$$\mu = p$$

Given the binary utilities defined in section 1, China's expected utility is given by:

$$U_C(A|\mu) = \mu,$$

$$U_C(P|\mu) = 1 - \mu.$$

China prefers A to P when, and only when:

$$\mu > 1 - \mu$$

$$\mu > \frac{1}{2}$$

China's critical belief threshold, therefore, is:

$$\hat{\mu} \in (0, 1) = \frac{1}{2}.$$

This belief threshold is determined by the utility function. If falling behind were more costly than the risks of accelerating, then the belief threshold would decrease: it would take more compelling evidence of stagnation for China to pause rather than accelerate.

Given that $p = 0.7$, China accelerates; the US obtains zero utility, and China receives a payoff with respect to the true state of the United States' frontier capabilities.

Hence, absent information design, the race is on: China accelerates whenever its prior that the United States is developing dangerous novel frontier capabilities is sufficiently high.

I turn now to solving the information design problem for the binary model.

2 Solving the Binary Game

In line with the standard Bayesian Persuasion result, the United States can increase its expected payoff by designing an information sharing regime that manipulates China's posterior beliefs.

Now, **U** designs and implements an eval – a statistical test of frontier model capabilities with error rates:

$$\alpha = P[s = 1|\theta = 1],$$

$$\beta = P[s = 1|\theta = 0].$$

U gains its maximal payoff whenever **C** pauses and 0 whenever **C** accelerates, so **U** attempts to minimize the probability that **C** accelerates.

Once China observes the signal from the eval, it updates its probability that the true capability is $\theta = 1$. Formally:

$$\mu_s = P(\theta = 1|s).$$

Because there are only two possible signals, there are two distinct posteriors.

$$\mu_1 \text{ – the belief after seeing the signal } s = 1.$$

$$\mu_0 \text{ – the belief after seeing the signal } s = 0.$$

China's posterior is given by:

$$\mu_1 = \frac{p\alpha}{p\alpha + (1-p)\beta}$$

$$\mu_0 = \frac{p(1-\alpha)}{p(1-\alpha) + (1-p)(1-\beta)}$$

Given the prior $p = 0.7$, the best **U** can do is to:

1. Keep $\mu_0 < \frac{1}{2}$ so that acceleration only happens after the signal $s = 1$ ("new dangerous capabilities");
2. Minimize $P(s = 1) = p\alpha + (1 - p)\beta$ subject to the previous constraint.

Lemma 1 (false positives never help).

> For any (α, β) with $\beta > 0$ there exists $(\alpha, 0)$ that weakly reduces $P(s = 1)$ while strictly decreasing μ_0 and weakly increasing μ_1. Hence an optimal test always sets $\beta* = 0$.
>
> **Proof.** Decrease β until it hits 0; both posteriors move in directions favorable to the sender while $P(S = 1)$ falls.

With $\beta = 0$, the first constraint becomes:

$$\mu_0 = \frac{p(1-\alpha)}{1 - p\alpha} \leq \frac{1}{2}.$$

$$\alpha \geq \frac{2p - 1}{p}$$

Lemma 2 (tight constraint).

> Given $\beta = 0$, $P(s = 1) = p\alpha$ is strictly increasing in α. The optimal choice, therefore, is the minimum feasible $\alpha*$:
>
> $$\alpha* = \frac{2p - 1}{p}$$

By Lemma 1 and Lemma 2, the optimal binary experiment is:

$$\alpha* = 0.57$$

$$\beta* = 0.$$

C only accelerates when it sees $\theta = 1$, and so **U**'s expected payoff rises to:

$$1 - P(s = 1) = 2(1 - p)$$

$$= 0.6$$

No other eval yields a higher payoff.

This result depends crucially on the assumption that the default scenario involves China taking a “bad” action from the perspective of the United States. If China believed that the probability of novel dangerous capabilities was low, such that it would pause by default, then **U**’s information design is not beneficial.

When that condition holds, however, **U** can design evals and share their results in order to make a pause more likely – thus contributing to stabilizing arms race dynamics.

3 Extension to a Continuous Model

I now enrich the model by treating the capability level and the signal as scalars.

Capability θ is drawn from a continuously–supported c.d.f. F on $[0,1]$. Denote the prior mean $m = E[\theta] \in (0,1)$.

U chooses an eval as an experiment – a family of conditional densities $g_\theta(.)_{\theta \in [0,1]}$ – and honestly commits to sharing the realized signal $s \in [0,1]$.

As before, Nature determines θ and **U** elicits a signal from its eval to share with China.

After observing s, China forms the posterior

$$\mu(s) = P(\theta > 0|s)$$

and plays P if $\mu(s) < \frac{1}{2}$ and A if $\mu(s) > \frac{1}{2}$.[2]

U’s payoff based on an experiment g_θ is

$$U_U(g_\theta) = 1 - P(\mu(s) > 1/2)$$

[2] Again, the belief threshold on the right hand side is a function of China’s utility function.

so it wants to minimize the probability mass of signals that push **C**'s posterior above the $\frac{1}{2}$ threshold.

Let π be the distribution of posteriors $\mu(s)$ induced by g_θ. Bayes' rule imposes the constraint:

$$\int_{[0,1]} \mu d\pi(\mu) = m$$

Given **U**'s payoff and the Bayesian constraint, **U**'s utility is concave in π insofar as it rewards weight below μ. By an extreme point argument, an optimum places weight on at most two posterior means (δ_1, where the posterior belief $> \frac{1}{2}$ and δ_0, where the posterior belief $\leq \frac{1}{2}$):

$$\pi^* = (1 - \gamma^*)\delta_0 + \gamma^* \delta_1$$

$$0 \leq \mu_0^* \leq \frac{1}{2} < \mu_1^* \leq 1$$

Where γ^* is the probability weight on the $\mu > \frac{1}{2}$ region. Solving Bayes' constraint yields

$$\gamma^* = \frac{m - \mu_0^*}{\mu_1^* - \mu_0^*}$$

Plugging in $\mu_0^* = \frac{1}{2}$ and $\mu_1^* = 1$ gives:

$$\gamma^* = \max\{0, 2m - 1\}$$

$$U_U^* = 1 - \gamma^* = \min\{1, 2(1 - m)\}$$

This game collapses into the fully binary model when we restrict θ to $\{0, 1\}$, with identical results when $m = p = 0.7$.

TODO – comment on how one could implement an eval with this sort of tuned probability distribution

4 Voluntary Disclosure

The Bayesian Persuasion optimum involves a process that would likely be unfamiliar to states: tuning the statistical properties of evals to induce strategic responses. In practice, states are instead likely to perform a given suite of evals and then strategically share a subset of their results. When the resulting evidence is "hard" – insofar as it cannot be false – Ali et al. (2024) demonstrate that voluntary disclosure can approximate the optimal information design equilibrium.

In the voluntary disclosure game, Nature again draws $\theta \in [0,1]$. The sender observes θ and may disclose any verifiable message $m \subseteq [0,1]$ containing the true state. Lying is impossible (θ must indeed lie in m), but silence – $m = [0,1]$ – is always feasible.

Utilities are given as in the previous models. After observing m, China updates its posterior

$$\mu(m) = E[\theta | \theta \in m]$$

and plays A if $\mu(m) > \frac{1}{2}$, P otherwise.

Let π^* be the two-posterior distribution from section 3:

$$\pi^* = (1-\gamma^*)\delta_0 + \gamma^*\delta_1$$

$$\gamma^* = \max\{0, 2m-1\}$$

By Lemma 2 of Ali et al. (2024), there is a finite partitional segmentation of observations of θ that achieves payoffs for $\mathbf{U}$ within ε of the Bayesian Persuasion optimum. The finite partitional segmentation assigns a message to different observed values of θ.

In our model, the optimal partition splits $[0,1]$ into three segments: M_0, M_1, and M_{rest}. $\mathbf{U}$ sends its message based on the partition into which its observation of θ falls.

$\theta \in M_1$ is disclosed as $m = M_1$;

$\theta \in M_0$ is disclosed as $m = M_0$;

$\theta \in M_{\text{rest}}$ yields the silent message $m = [0,1]$.

China accelerates if, and only if, $m = M_1$.

It can be shown that

$$\Pr(\theta \in M_1) = \gamma^\star \pm \varepsilon$$

such that U_U is within ε of the Bayesian Persuasion optimum.

TODO – I think I still haven't grasped exactly how this proof works. I want to try to work through a proof in the body of the text, to try to make the results more intuitive for an unfamiliar reader (with a relative lack of mathematical abstraction/generality, based on the fact that we're working through a concrete case).

Hence, in the absence of a binding commitment mechanism, verifiable evals can combine with a partitioned information sharing regime to let the Untied States capture the benefits of strategic transparency.

Conclusion

TODO

References

S. Nageeb Ali, Andreas Kleiner, and Kun Zhang. From design to disclosure, 2024. URL https://arxiv.org/abs/2411.03608.

Stuart Armstrong, Nick Bostrom, and Carl Shulman. Racing to the precipice: A model of artificial intelligence development. *AI and Society*, 31(2):201–206, 2016. doi: 10.1007/s00146-015-0590-y.

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

Nicholas Emery-Xu, Andrew Park, and Robert Trager. Uncertainty, information, and risk in international technology races. *Journal of Conflict Resolution*, 68(10):2019–2047, 2024. doi: 10.1177/00220027231214996. URL https://journals.sagepub.com/doi/abs/10.1177/00220027231214996.

Emir Kamenica and Matthew Gentzkow. Bayesian persuasion. *American Economic Review*, 2011.

Eoghan Stafford, Robert F. Trager, and Allan Dafoe. Safety not guaranteed: International races for risky technologies. 2022.