Reputation, Verification, and Strategic Deception

Noah Fehr

December 13, 2025

Abstract

I develop a dynamic signaling model motivated by China's use of strategic ambiguity in the Taiwan Strait. In this model, an authoritarian sender can misrepresent its intentions at a fixed frequency. The sender trades off short-term gains from deception against the risk that periodic monitoring will expose a lie and undermine its credibility. I characterize the perfect Bayesian equilibrium of this reputation game and show that the deceptive type chooses an optimal lying frequency x^* that balances immediate benefits and long-run reputational costs. The equilibrium lying rate falls when monitoring becomes more intensive or when players are more patient, and it rises when the short-run payoff from deception is larger. Across parameter values, the model produces outcomes that range from full honesty to full deception, along with intermediate cases in which the sender mixes between truth-telling and lying. These results provide a framework for understanding strategic ambiguity and the role of verification in international political conflicts.

1 Motivation

1.1 Signaling in the Taiwan Strait

For decades, China's behavior in the Taiwan Strait has featured ambiguous signaling and calculated posturing, creating uncertainty for geopolitical rivals such as the US and the Republic of China (the national government of Taiwan). Beijing refuses to clarify the exact conditions that would prompt the use of force against Taiwan, while reiterating that the People's Republic of China will not tolerate Taiwan's independence. For example, the Anti-Secession Law states that major incidents promoting secession or the exhaustion of peaceful unification efforts would trigger military action, yet even these conditions remain vague (Tiong & Hoo, 2024). This ambiguity gives China flexibility and weakens its adversaries' ability to predict its next step.

While the Taiwan Strait has increasingly defined geopolitical tensions in recent years, Beijing's use of ambiguity is not new. In the 1995–1996 Third Taiwan Strait Crisis, China postured with missile tests and amphibious exercises, but ultimately did not invade Taiwan. Although US intelligence correctly understood these moves as a show of force rather than wartime preparations, the signals effectively communicated dissatisfaction and a willingness to challenge perceived pro-independence moves (Gunness & Saunders, 2022). Similar dynamics re-emerged in 2022 with Beijing's live-fire drills and temporary blockade following Speaker Pelosi's visit to Taipei (Tiong & Hoo, 2024). China uses military exercises, patrols, and sharp rhetoric to signal its capabilities and resolve, while still preserving deniability and room to shift course. As a result, the US and Taiwan must continually evaluate whether such signals represent credible commitments, political theater, or something in between.

The formal model that follows abstracts from these specific military moves and diplomatic episodes but keeps three core features of the Taiwan setting. First, China repeatedly issues public claims about its intentions and capabilities that can be truthful or not. Second, the United States and other observers only occasionally gain access to hard verification through satellite imagery, intelligence, or third-party monitoring, which sometimes reveals whether a claim was false. Third, reputational assessments are sticky but fragile: a history of seemingly honest behavior makes later signals more persuasive, yet a single verifiable deception can sharply reduce the credibility of subsequent statements; in our model, this erases all future credibility. The model distills these ingredients into a simple dynamic game in which a sender chooses a lying frequency, anticipating that future monitoring and belief updating will feed back into its ability to shape the receiver's behavior.

1.2 Informational Control and Reputational Consequences

The model focuses on China's authoritarian regime and its resulting ability to shape information flows in ways that are generally unavailable to states with a free press. The Chinese Communist Party (CCP) maintains strict control over domestic media and, by extension, exerts substantial influence over how its behavior is portrayed internationally. Beyond downplaying incidents at home, China can saturate parts of the global information space with its preferred narratives, complicating efforts by outsiders to verify claims. The US State Department estimates that Beijing spends billions of dollars annually on "deceptive and coercive methods" to bend the global information environment in its favor (Al Jazeera, 2023). Such efforts work to bias international perceptions and to induce responses that align more closely with Beijing's interests.

Because of this information control, China can deceive or selectively release information with a lower risk of immediate exposure than many of its democratic rivals; the risk that a false statement is quickly uncovered by independent media or institutions is higher for Western governments than for China. A salient example is President Xi Jinping's repeated assurances in 2015 that China did not "intend to pursue militarization" of the artificial islands of the Spratlys (Panda, 2016). These assurances were intended to defuse international concern while China continued to build airstrips and facilities it insisted were "dual-use." China's prior reputation and the difficulty of real-time verification lent these statements initial plausibility and helped dampen the international community's resolve to intervene. Subsequent satellite imagery in late 2016, however, exposed anti-aircraft guns and other military installations, showing that Xi's pledge was misleading (Panda, 2016).

This is the core logic of the model: a sender issues statements that are initially taken seriously, but once verification reveals deception, credibility collapses. In our framework, a verifiably false claim functions as a checked signal that exposes a lie and leads the receiver to ignore the sender's future statements. The Spratlys incident provides a clear example of this dynamic. While China did not face major domestic costs, its shifting explanations generated international skepticism about its assurances (Panda, 2016), suggesting movement toward a reputational cliff.

A similar logic operates in the Taiwan context, which can be viewed as a new instance of the same game. China's signaling is directed not only at the United States but also at the Taiwanese public; Taiwan is frequently described as the "front line" of China's information campaign. After the Russian invasion of Ukraine, Chinese state media pushed narratives portraying the US as unreliable, and surveys subsequently recorded roughly a 10% decline in Taiwanese trust in the United States (Al Jazeera, 2023).

For tractability, the model treats the United States as the sole receiver, but the Taiwan case underscores how this reputational mechanism naturally extends to multiple receivers, each updating beliefs in response to China's statements and their eventual verification.

1.3 Reputational Dynamics

Repeated signaling feeds into the model's reputational calculus. In an iterative rivalry like that between the US and China, individual statements and episodes do not stand alone; they accumulate into a history that shapes how future claims are interpreted. In the model, the United States, as the primary observer, holds a belief μ_t about whether China is an honest or deceptive type and updates this belief based on China's statements and the subset of outcomes that become verifiable.

The key simplifying assumption is that a detected lie is catastrophic for China's reputation. When a checked claim is observed to be truthful, the US updates its belief according to Bayes' rule, raising μ and making future statements more persuasive. By contrast, when a checked claim is revealed to be false, the posterior collapses to

$$\mu_{t+1} = 0,$$

which in the baseline model is an absorbing state.[1]Once a lie is caught, the US

[1] In reality, reputations may recover gradually after deception is exposed, through subsequent behavior or changes in context. The model takes a stark limit in which detection wipes out credibility in order to keep the analysis tractable and to highlight the underlying incentive mechanism. Because of this design choice, this model is best understood in the context of a specific conflict, like the Spratlys or Taiwan, as opposed to long-run US–China competition across many domains.

infers that China must be the deceptive type, and subsequent Chinese statements no longer move American beliefs or policy.

As long as no lie has been verifiably exposed, however, China still benefits from whatever reputation it has built. A higher μ_t enlarges its effective action space by making it easier to shift US expectations. Deception yields short-run gains through the immediate payoff $f(\mu_t)$, but it also increases the chance of triggering a checked falsehood that sends μ_t to zero. Selective honesty therefore matters in the model because it can reduce the probability of ever triggering the reputational collapse and improve future payoffs through a higher μ. The equilibrium lying frequency x^* studied in the next section is precisely the choice that balances the immediate benefits from manipulation against the discounted loss of all future influence once a lie is verifiably exposed.

In sum, the incomplete information, strategic signaling, and stark reputational consequences of relations between China and the US over Taiwan define the environment of the model. Historical and contemporary evidence suggests that China has at times leveraged ambiguity and deception, and at other times invested in credibility, to shape US and Taiwanese perceptions in both previous conflicts and the present tension over Taiwan itself.

2 Model

The model translates this environment into a simplified sender–receiver game. China is the sender, issuing public claims about its behavior or intentions; the United States is the receiver, forming beliefs about whether those claims are truthful and adjusting its policy stance accordingly. Each period corresponds to a round of signaling over some issue followed by a chance that independent monitoring reveals whether the claim was accurate. The monitoring parameter θ captures how often such reliable checks occur, while the belief μ_t summarizes the United States' current assessment of whether China is fundamentally an "honest" or "deceptive" type. In this way, the abstract objects in the model map directly onto the empirical setting described previously.

There are two players: China and the United States (US). Time is discrete and indexed by $t = 0, 1, 2, \ldots$.

2.1 Types, beliefs, and monitoring

At $t = 0$, Nature draws China's type $\tau \in \{H, D\}$, where H denotes an honest type and D denotes a deceptive type. The initial prior, or the probability that China is honest, is

$$\Pr(\tau = H) = \mu_0 \in (0, 1)$$

This prior is common knowledge, shared by both parties. At time t, the US holds a belief $\mu_t \in [0, 1]$ given by

$$\mu_t = \Pr(\tau = H \mid \text{history up to } t),$$

so μ_t summarizes the US's reputation-based assessment of China's type.

In each period t, after China makes a claim, the United States verifies that claim with exogenous probability $\theta \in (0, 1]$. We interpret θ as the fraction of interactions in which China's communication is *reliably checked.* If a check occurs, it perfectly reveals whether the observed claim is truthful. Checks are independent of the action itself.

2.2 Actions and lying frequency

China's type determines its feasible behavior:

- The honest type $\tau = H$ never lies: in every period and state, China reports truthfully.
- The deceptive type $\tau = D$ may misrepresent reality. We summarize China's behavior by a lying frequency $x \in [0, 1]$, where x is the probability (or long-run frequency) with which it sends a false claim in a given period.

We focus on stationary behavior by the deceptive type, so that the equilibrium x^* is constant across periods and known to both players.

2.3 Payoffs

Let μ_t denote US's belief at the beginning of period t. The current-period payoff to China from successfully deceiving the United States at belief μ_t is given by

$$f(\mu_t),$$

where $f : [0, 1] \to R$ is an increasing, non-negative function. Thus, the reputational benefit of being perceived as honest is higher when μ_t is larger. There is no immediate benefit to being honest.

Future payoffs enter through a value function $V : [0, 1] \to R$, which maps the belief μ about China's honesty into the continuation value of future interactions. We assume:

$$V'(\mu) > 0 \quad \text{and} \quad V''(\mu) < 0,$$

so that reputation is valuable (higher μ is better) but with diminishing marginal returns. We also assume that V is invertible on its range. Intertemporal tradeoffs are captured by a common discount factor

$$\delta \in (0, 1),$$

which is common knowledge.

2.4 Beliefs and Bayesian updating

Given a belief $\mu_t \in [0, 1]$ and a lying frequency $x \in [0, 1]$ for the deceptive type, we define the Bayesian updating rule for beliefs following an observed truthful claim.

Let "Truth" denote the event that China's claim is truthful in a period where the US checks. Under the honest type, this event occurs with probability 1. Under the deceptive type, it occurs with probability $1 - x$. Hence,

$$\Pr(\text{Truth} \mid \tau = H) = 1, \qquad \Pr(\text{Truth} \mid \tau = D) = 1 - x,$$

and Bayes' rule implies that the posterior probability that China is honest after observing a truthful claim is

$$\begin{aligned} \mu_{t+1} &:= \Pr(\tau = H \mid \text{Truth}) && (1) \\ &= \frac{\Pr(\text{Truth} \mid \tau = H)\Pr(\tau = H)}{\Pr(\text{Truth})} && (2) \\ &= \frac{1 \cdot \mu_t}{\mu_t \cdot 1 + (1 - \mu_t)(1 - x)} = \frac{\mu_t}{1 - x + x\mu_t}. && (3) \end{aligned}$$

μ_{t+1} thus describes how the US's belief about China's type evolves when a checked claim is found to be truthful, given the deceptive type's lying frequency x.

In periods where the US either does not check $(1 - \theta)$ or observes a lie, beliefs are updated accordingly: μ_{t+1} falls to 0 if the US catches a lie and remains equal to μ_t if no check is performed. In this sense, the messages sent in no-check periods are treated as cheap talk and do not affect beliefs on the equilibrium path, reflecting the explicitly conflicting interests of the US and China (Crawford & Sobel, 1982).

In summary, the belief μ_{t+1} evolves as follows:

1. Check reveals truth: Bayesian update as shown above
2. Check reveals lie: μ_{t+1} drops to 0
3. No check performed: $\mu_{t+1} = \mu_t$

Note that throughout the proofs, when the value μ_{t+1} is referenced, I am referring to Case (1), the Bayesian update. When μ_{t+1} evaluates to Case (2) or Case (3), I substitute the corresponding values in for μ_{t+1}.

2.5 Dynamic game and equilibrium

The interaction between China and the US can be viewed as a repeated Bayesian game. At the start of each period t:

1. The public belief μ_t about China's type is given.
2. China (of type τ) issues a claim; if $\tau = D$, her behavior is summarized by a lying frequency $x \in [0, 1]$.
3. With probability θ, the United States checks the claim and observes whether it is truthful; otherwise no verification occurs.

4. Payoffs from the present period are realized, and beliefs update as described in 2.4. The next-period belief μ_{t+1} determines the continuation value via $V(\mu_{t+1})$.

China's objective is to choose a lying frequency x to maximize the expected discounted sum of current deception payoffs $f(\mu_t)$ and discounted future reputation payoffs $\delta V(\mu_{t+1})$, given the monitoring intensity θ and the belief update rule. In subsequent sections, we characterize stationary equilibria in which the deceptive type chooses a constant $x^* \in [0, 1]$ and beliefs evolve according to 2.4.

3 Equilibrium

3.1 Strategies, beliefs, and equilibrium concept

A stationary strategy for China specifies, for each type $\tau \in \{H, D\}$ and belief $\mu \in [0, 1]$, the lying frequency $x(\tau, \mu) \in [0, 1]$. By construction, the honest type never lies, so

$$x(H, \mu) = 0 \quad \text{for all } \mu,$$

while the deceptive type chooses a stationary lying frequency $x(D, \mu) \equiv x \in [0, 1]$. The United States observes China's claim and, whenever it checks, updates its belief about China's type according to Bayes' rule. Therefore, in this model, the receiver's strategy is trivial (since monitoring is exogenous), so the US' role in the PBE is captured entirely by the belief-update rule.

Let σ denote China's strategy and μ a belief system that assigns, after every history, a belief over types summarized by $\mu_t \in [0, 1]$.

Assessment: An assessment is a pair (σ, μ) where σ is a strategy profile (China's strategy) and μ is a belief system derived from some prior over types and observed histories (US's prior about China).

Perfect Bayesian equilibrium: An assessment (σ^*, μ^*) is a perfect Bayesian equilibrium (PBE) if:

1. **Sequential rationality.** Given beliefs μ^*, each type of China chooses a lying frequency that maximizes its expected discounted payoff, while the United States updates its beliefs according to Bayes' rule, with checks occurring exogenously with probability θ.

2. **Belief consistency.** On the equilibrium path, beliefs are updated by Bayes' rule whenever possible. In particular, following any checked truthful claim, beliefs are updated using

$$\mu_{t+1} = \frac{\mu_t}{1 - x + x\mu_t}.$$

   Off the equilibrium path, μ^* is derived from σ^* and the prior in a Bayesian-consistent way whenever Bayes' rule is applicable.

We focus on stationary PBEs in which the deceptive type chooses a constant $x^* \in [0, 1]$ and beliefs evolve according to Bayes' rule whenever a checked truthful claim is observed.

3.2 Payoffs and the deceptive type's problem

Fix a belief $\mu_t \in (0,1)$ at the start of a period and a stationary lying frequency $x \in [0,1]$ for the deceptive type. Recall:

- $f(\mu_t)$ is the immediate payoff from successfully deceiving when the belief is μ_t.
- $V(\mu_t)$ is the continuation value of future interactions when the reputation is μ_t, with $V'(\mu_t) > 0$ and $V''(\mu_t) < 0$. [2]
- With probability θ, the United States checks the claim; with probability $1-\theta$ it does not.

If the deceptive type lies in the current period and is caught when checked, the continuation belief is $\mu_{t+1} = 0$, so the continuation payoff is $V(0)$. We can write the deceptive type's payoff from lying as

$$L(\mu_t) = f(\mu_t) + \delta(1-\theta)V(\mu_t) + \delta\theta V(0). \tag{4}$$

If the deceptive type behaves honestly in the current period, there is no immediate payoff gain (by assumption), but a checked truthful claim leads to the posterior μ_{t+1} updated via the Bayesian update and continuation value $V(\mu_{t+1})$. Hence the payoff from honesty at lying frequency x is

$$H(x;\mu_t) = \delta\theta V(\mu_{t+1}) + \delta(1-\theta)V(\mu_t). \tag{5}$$

Given a stationary lying frequency x, the deceptive type's expected payoff in the current period is

$$U(x;\mu_t) = xL(\mu_t) + (1-x)H(x;\mu_t). \tag{6}$$

Assumption 1 (Concavity of the deceptive type's payoff)**.** *For each belief $\mu \in (0,1)$, the deceptive type's payoff $U(x;\mu)$ is strictly concave in the lying frequency x on $[0,1]$. This property holds under mild conditions on the payoff primitives. The function $U(x;\mu)$ combines an immediate payoff $f(\mu)$ that enters linearly in x with a continuation value $\delta\theta V(\mu_{t+1})$ that follows a checked truthful claim, where the updated belief $\mu_{t+1} = \frac{\mu_t}{(1-x+x\mu_t)}$ is increasing and smooth in x on $(0,1)$. If $f(\mu)$ grows at most linearly in μ and the reputation value $V(\mu)$ is sufficiently concave, so that $V''(\mu) < 0$ and marginal returns to reputation diminish, then the composition $V(\mu')$ is concave in x. Under these conditions the second derivative $U''(x;\mu)$ is negative. The intuition is that as x becomes larger, the risk of a reputational collapse increases faster than the growth of potential lying payoffs, thus reducing the net marginal benefit of additional deception.*

[2] In a fully specified dynamic game, $V(\mu)$ would be obtained as the discounted sum of future payoffs conditional on belief μ. For tractability, I treat $V(\mu)$ as a reduced-form continuation value rather than deriving it explicitly. It can be interpreted as the present value of future interactions that would arise in an infinite-horizon extension of the stage game.

These conditions ensure that the reputational component dominates any curvature in the immediate payoff. To avoid imposing stronger structure on f and V, I take $U''(x;\mu) < 0$ as a regularity condition that guarantees a unique interior optimum whenever one exists.

The deceptive type chooses x to maximize $U(x;\mu)$.

Lemma 1 (Derivative of the deceptive type's payoff)**.** *$U(x;\mu_t)$ is differentiable on $(0,1)$ and*

$$U'(x;\mu_t) = L(\mu_t) - H(x;\mu_t) + (1-x)H_x(x;\mu_t), \tag{7}$$

where

$$H_x(x;\mu_t) = \delta\theta V'(\mu_{t+1})\,\frac{\partial \mu_{t+1}}{\partial x}. \tag{8}$$

Proof. Differentiating (6) with respect to x yields

$$U'(x;\mu_t) = L(\mu_t) - H(x;\mu_t) + (1-x)H_x(x;\mu_t),$$

and differentiating (5) with respect to x gives

$$H_x(x;\mu_t) = \delta\theta V'(\mu_{t+1})\,\frac{\partial \mu_{t+1}}{\partial x}.$$

This establishes (7) and (8). □

A stationary optimal lying frequency $x^* \in [0,1]$ is a maximizer of $U(x;\mu)$ on $[0,1]$. Any interior solution $x^* \in (0,1)$ must satisfy the first-order condition

$$U'(x^*;\mu) = 0,$$

while boundary solutions $x^* \in \{0,1\}$ arise when $U(\cdot;\mu)$ is maximized at the endpoints.

3.3 Pure-strategy equilibria

We first characterize the conditions under which the deceptive type chooses $x = 0$ (never lies) or $x = 1$ (always lies) in equilibrium.

3.3.1 Honesty equilibrium: $x^* = 0$

When $x = 0$, we have $\mu_{t+1} = \mu_t$ and $\frac{\partial \mu_{t+1}}{\partial x} = \mu_t(1-\mu_t)$. Substituting into (5) and (8), we obtain $H(0;\mu_t) = \delta V(\mu_t)$ and

$$H_x(0;\mu_t) = \delta\theta V'(\mu_t)\,\mu_t(1-\mu_t).$$

From Lemma 1,

$$\begin{aligned} U'(0;\mu_t) &= L(\mu_t) - H(0;\mu_t) + H_x(0;\mu_t) && (9)\\ &= f(\mu_t) + \delta(1-\theta)V(\mu_t) + \delta\theta V(0) - \delta V(\mu_t) + \delta\theta V'(\mu_t)\mu_t(1-\mu_t) && (10)\\ &= f(\mu_t) + \delta\theta\Big[V(0) - V(\mu_t) + V'(\mu_t)\frac{\mu_t}{1-\mu_t}\Big]. && (11) \end{aligned}$$

Proposition 1 (Honesty equilibrium)**.** *If*

$$f(\mu_t) + \delta\theta\Big[V(0) + V'(\mu_t)\frac{\mu_t}{1-\mu_t}\Big] \leq \delta\theta V(\mu_t), \tag{12}$$

then there exists a stationary PBE in which the deceptive type chooses $x^* = 0$ and China never lies on the equilibrium path.

Proof. Condition (12) is equivalent to $U'(0;\mu) \leq 0$ after rearranging (11). If $U'(0;\mu) \leq 0$, then for x in a neighborhood of 0, increasing x cannot raise the deceptive type's payoff. Because $x \in [0,1]$ and U is continuous and concave in the relevant region (given $V'' < 0$), the best response is $x^* = 0$. Given $x^* = 0$, beliefs follow from Bayes' rule, so the resulting assessment is a stationary PBE. □

Intuitively, the right-hand side of (12), $\delta\theta V(\mu_t)$, is the discounted continuation value from preserving reputation by remaining honest. The left-hand side is the sum of the immediate gain from lying, $f(\mu)$, the continuation value after being caught ($\delta\theta V(0)$), and the marginal reputational loss captured by $\delta\theta V'(\mu_t)\frac{\mu_t}{1-\mu_t}$. When the long-run value of credibility outweighs these gains, the deceptive type optimally chooses $x^* = 0$.

3.3.2 Full deception equilibrium: $x^* = 1$

When $x = 1$, a checked truthful claim occurs with probability 0 for the deceptive type. Nevertheless, conditional on an (off-path) checked truthful claim, Bayes' rule implies

$$\mu_{t+1} = \frac{\mu_t}{1 - 1 + 1 \cdot \mu_t} = 1,$$

and hence, from (5), $H(1;\mu_t) = \delta V(1)$. Evaluating (7) at $x = 1$ yields

$$\begin{aligned} U'(1;\mu_t) &= L(\mu_t) - H(1;\mu_t) && (13)\\ &= f(\mu_t) + \delta(1-\theta)V(\mu_t) + \delta\theta V(0) - \delta V(1) && (14)\\ &= f(\mu_t) + \delta\theta\big[V(0) - V(1)\big]. && (15) \end{aligned}$$

Proposition 2 (Full deception equilibrium)**.** *If*

$$f(\mu_t) + \delta\theta V(0) \geq \delta\theta V(1), \tag{16}$$

then there exists a stationary PBE in which the deceptive type chooses $x^* = 1$ and always lies on the equilibrium path.

Proof. Condition (16) is equivalent to $U'(1;\mu) \geq 0$ using (15). If $U'(1;\mu) \geq 0$, then reducing x from 1 locally does not increase the deceptive type's payoff. By continuity of U, the best response is $x^* = 1$. Given $x^* = 1$, beliefs again follow from Bayes' rule when checks occur, and the resulting assessment is a stationary PBE. □

Here the left-hand side of (16) is the expected payoff from lying (immediate payoff $f(\mu)$ plus continuation after being caught, $\delta\theta V(0)$), while the right-hand side is the discounted value of maintaining a fully honest reputation $V(1)$. When the former exceeds the latter, the deceptive type chooses $x^* = 1$ and lying is a best response.

3.4 Mixed-strategy equilibrium

We now consider the interior case $x^* \in (0,1)$, where neither pure strategy condition holds. This requires

$$\delta\theta\Big[V(\mu_t) - V'(\mu_t)\frac{\mu_t}{1-\mu_t}\Big] < f(\mu_t) + \delta\theta V(0) < \delta\theta V(1). \tag{17}$$

The left inequality is the failure of the honesty condition (12), while the right inequality is the failure of the full deception condition (16), ensuring $U'(0) > 0$ and $U'(1) < 0$. In this parameter region, the optimal lying frequency x^* is interior and satisfies $U'(x^*; \mu) = 0$.

From Lemma 1 and the expressions for L and H, the first-order condition for an interior optimum can be written as

$$f(\mu_t) + \delta\theta\big[V(0) - V(\mu_{t+1})\big] + \delta\theta(1-x)V'(\mu_{t+1})\,\frac{\partial\mu_{t+1}}{\partial x} = 0. \tag{18}$$

For analysis, it is convenient to reparametrize in terms of the posterior $y = \mu_{t+1}$. Recall that in the case China is found to be honest:

$$\mu_{t+1} = \frac{\mu_t}{1 - x(1-\mu_t)}, \qquad \frac{\partial\mu_{t+1}}{\partial x} = \frac{(1-\mu_t)y^2}{\mu_t},$$

and

$$1 - x = \frac{\mu_t(1-y)}{(1-\mu)y}, \qquad (1-x)\frac{\partial\mu_{t+1}}{\partial x} = y(1-y).$$

Substituting these expressions into (18) yields

$$f(\mu_t) + \delta\theta\big[V(0) - V(y)\big] + \delta\theta V'(y)y(1-y) = 0. \tag{19}$$

Rearranging,

$$V(y) - y(1-y)V'(y) = V(0) + \frac{f(\mu_t)}{\delta\theta}. \tag{20}$$

Define

$$\phi(y) := V(y) - y(1-y)V'(y).$$

Then the interior equilibrium posterior y^* solves

$$\phi(y^*) = V(0) + \frac{f(\mu_t)}{\delta\theta}. \tag{21}$$

Given y^*, the corresponding lying frequency is obtained from the identity

$$y^* = \mu_{t+1} = \frac{\mu_t}{1 - x(1-\mu_t)},$$

which yields

$$x^* = \frac{1 - \frac{\mu_t}{y^*}}{1 - \mu_t}. \tag{22}$$

Proposition 3 (Interior equilibrium)**.** *Suppose* (17) *holds and* ϕ *is strictly increasing on* $(0, 1)$*. Then there exists a unique* $y^* \in (0, 1)$ *solving* (21)*, and the corresponding lying frequency* x^* *given by* (22) *lies in* $(0, 1)$*. Under Assumption 1, this* x^* *is the unique optimal lying frequency for the deceptive type, and the assessment induced by* x^* *together with belief updating via* μ' *constitutes a stationary PBE.*

Proof. Since ϕ is continuous and strictly increasing on $(0, 1)$, the equation (21) has at most one solution in this interval. The mixed-strategy conditions in (17) ensure that the right-hand side of (21) lies strictly within the range of ϕ on $(0, 1)$, which implies that a unique $y^* \in (0, 1)$ exists. The mapping from y^* to x^* in (22) is one-to-one on $(0, 1)$, so the corresponding x^* is also unique and satisfies $U'(x^*; \mu) = 0$.

Assumption 1 states that $U(x; \mu)$ is strictly concave in x, which guarantees that any solution to $U'(x; \mu) = 0$ is the unique global maximizer of $U(\cdot; \mu)$ on $[0, 1]$. Thus the deceptive type has a unique interior best response in this region. Since beliefs update according to Bayes' rule via the formula for μ', the resulting assessment defines a stationary PBE. □

3.5 Comparative statics

We now examine how the interior equilibrium x^* responds to changes in the key parameters. Let y^* and x^* be defined by (21)–(22).

Proposition 4 (Comparative statics)**.** *Assume the interior equilibrium in Proposition 3 exists and* $\phi'(y^*) > 0$*. Then:*

1. ***Monitoring and patience.*** *The equilibrium lying frequency decreases in the monitoring intensity* θ *and the discount factor* δ*:*

$$\frac{dx^*}{d\theta} < 0, \qquad \frac{dx^*}{d\delta} < 0.$$

2. ***Reputation (belief) effects.*** *The effect of* μ *on* x^* *is ambiguous:*

$$\frac{dx^*}{d\mu} = \underbrace{\frac{\partial x}{\partial y}}_{\frac{\mu}{(1-\mu)(y^*)^2}} \frac{dy^*}{d\mu} + \underbrace{\frac{\partial x}{\partial \mu}}_{-\frac{1-y^*}{y^*(1-\mu)^2}},$$

where the first term is increasing in $f'(\mu)$ *while the second term is negative. When* $f'(\mu)$ *is large, the temptation effect of a better reputation can dominate the mechanical reputational cost effect.*

Although the sign of $\frac{\partial x^*}{\partial \mu}$ *cannot be determined in general, the structure of the incentives suggests that* $x^*(\mu)$ *may exhibit non-monotonic behavior.*

When μ is small, a marginal increase in reputation raises the immediate payoff from deception through a higher value of $f(\mu)$, which can increase x^*. For larger values of μ, the continuation value $V(\mu)$ becomes more important, and the cost of jeopardizing a valuable reputation becomes dominant, which can reduce x^*. These two forces imply that $x^*(\mu)$ may be hump shaped: increasing for low values of μ and decreasing once reputation becomes sufficiently valuable. This pattern is most likely when $f(\mu)$ increases rapidly for intermediate values of μ and when $V(\mu)$ is strongly concave. The exact shape of $x^*(\mu)$ ultimately depends on the specific functional forms of f and V, so a monotonicity result cannot be established without additional assumptions.

Proof. Differentiate (20) implicitly. For a generic parameter $\kappa \in \{\mu, \theta, \delta\}$,

$$\phi'(y^*)\frac{dy^*}{d\kappa} = \frac{\partial}{\partial\kappa}\Big(V(0) + \frac{f(\mu)}{\delta\theta}\Big).$$

For $\kappa = \theta$,

$$\frac{dy^*}{d\theta} = \frac{-f(\mu)}{\delta\theta^2\,\phi'(y^*)}.$$

For $\kappa = \delta$,

$$\frac{dy^*}{d\delta} = \frac{-f(\mu)}{\delta^2\theta\,\phi'(y^*)}.$$

Using $\frac{\partial x}{\partial y} = \frac{\mu}{(1-\mu)(y^*)^2} > 0$ and $f(\mu) \geq 0$, $\phi'(y^*) > 0$, we obtain

$$\frac{dx^*}{d\theta} = \frac{\partial x}{\partial y}\frac{dy^*}{d\theta} = \frac{\mu}{(1-\mu)(y^*)^2}\cdot\frac{-f(\mu)}{\delta\theta^2\,\phi'(y^*)} < 0,$$

and similarly

$$\frac{dx^*}{d\delta} = \frac{\partial x}{\partial y}\frac{dy^*}{d\delta} = \frac{\mu}{(1-\mu)(y^*)^2}\cdot\frac{-f(\mu)}{\delta^2\theta\,\phi'(y^*)} < 0.$$

For $\kappa = \mu$, we use

$$\frac{dy^*}{d\mu} = \frac{f'(\mu)}{\delta\theta\,\phi'(y^*)},$$

and the decomposition

$$\frac{dx^*}{d\mu} = \frac{\partial x}{\partial y}\frac{dy^*}{d\mu} + \frac{\partial x}{\partial\mu} = \frac{\mu}{(1-\mu)(y^*)^2}\cdot\frac{f'(\mu)}{\delta\theta\,\phi'(y^*)} - \frac{1-y^*}{y^*(1-\mu)^2}.$$

The second term is negative, while the first term is positive when $f'(\mu) > 0$; hence the overall sign is ambiguous and depends on the relative magnitude of the temptation effect (via $f'(\mu)$) and the reputational cost effect (via V', V'' encoded in $\phi'(y^*)$). □

Taken together, Propositions 1–4 summarize how the equilibrium lying frequency x^* responds to changes in monitoring intensity, patience, reputation, and the immediate payoff from deception.

Beyond the lying frequency itself, it is useful to consider two related equilibrium objects. First, the deceptive type's expected payoff in equilibrium is given by $U(x^*; \mu_t)$, which is the value of its objective when it chooses the optimal lying frequency x^*. Comparative statics for $U(x^*; \mu)$ mirror those for x^*: an increase in the monitoring intensity θ or in the discount factor δ tends to reduce the deceptive type's payoff, since these changes raise the value of preserving reputation and induce a lower equilibrium lying frequency.

Second, the ex ante probability that a lie is detected in a given period can be written as

$$\Pr(\text{caught lie}) = \theta\,(1 - \mu_t)\,x^*.$$

Monitoring affects this probability in two opposing ways. A larger value of θ raises the chance that a given lie is checked, but it also lowers x^* by making deception less attractive. The net effect is therefore muted, and the detection rate need not rise proportionally with monitoring. This highlights how changes in monitoring intensity can have limited marginal impact once strategic adjustments by the sender are taken into account.

4 Illustrative functional forms and equilibrium regions

The equilibrium characterization in Propositions 1–3 is fully general in the functions f and V. In this section, we introduce simple parametric forms for f and V and show how the conditions for honesty, mixed, and full-deception equilibria translate into graphical regions.

4.1 Parametric specifications for f and V

We consider a family of specifications in which the payoff from lying is scaled by a parameter $\alpha > 0$ and the reputation value V is strictly increasing and concave.

Lying payoff. Let

$$f(\mu; \alpha) = \alpha\,\tilde{f}(\mu),$$

where $\tilde{f} : [0,1] \to R_+$ is an increasing function of μ. A simple example is $\tilde{f}(\mu) = \mu$, which captures the idea that deception is more rewarding when the sender enjoys a stronger reputation. α governs the overall scale of immediate gains from deception relative to the prior, μ.

Reputation value. Let V be twice differentiable, strictly increasing, and concave. A convenient benchmark is a quadratic:

$$V(\mu) = v_0 + v_1\mu - \frac{\gamma}{2}\mu^2,$$

with $v_1 > 0$ and $\gamma > 0$. Then $V'(\mu) = v_1 - \gamma\mu > 0$ for μ in the relevant range and $V''(\mu) = -\gamma < 0$, matching our assumptions.

We fix (δ, θ, V) and study how the equilibrium x^* varies with (μ, α).

4.2 Thresholds and regions in (μ, α)-space

Recall the pure-strategy conditions from Propositions 1 and 2, now written in terms of $f(\mu; \alpha)$.

Honesty condition. The honesty equilibrium $x^* = 0$ is supported whenever

$$f(\mu; \alpha) + \delta\theta\Big[V(0) + V'(\mu)\frac{\mu}{1-\mu}\Big] \leq \delta\theta V(\mu). \tag{23}$$

For a given μ, this is an upper bound on α. Using $f(\mu; \alpha) = \alpha\,\tilde{f}(\mu)$, (23) is equivalent to

$$\alpha \leq \alpha_0(\mu) := \frac{\delta\theta}{\tilde{f}(\mu)}\Big[V(\mu) - V(0) - V'(\mu)\frac{\mu}{1-\mu}\Big]. \tag{24}$$

Full-deception condition. The full-deception equilibrium $x^* = 1$ is supported whenever

$$f(\mu; \alpha) + \delta\theta V(0) \geq \delta\theta V(1). \tag{25}$$

Again substituting $f(\mu; \alpha) = \alpha\,\tilde{f}(\mu)$ gives

$$\alpha \geq \alpha_1(\mu) := \frac{\delta\theta}{\tilde{f}(\mu)}\big[V(1) - V(0)\big]. \tag{26}$$

Classification of regions. For each belief $\mu \in (0, 1)$, the functions $\alpha_0(\mu)$ and $\alpha_1(\mu)$ define three regions in the (μ, α)-plane:

$$\begin{aligned}
\mathcal{R}^0 &= \big\{(\mu, \alpha) : \alpha \leq \alpha_0(\mu)\big\} && \text{(unique honesty equilibrium } x^* = 0);\\
\mathcal{R}^1 &= \big\{(\mu, \alpha) : \alpha \geq \alpha_1(\mu)\big\} && \text{(unique full-deception equilibrium } x^* = 1);\\
\mathcal{R}^m &= \big\{(\mu, \alpha) : \alpha_0(\mu) < \alpha < \alpha_1(\mu)\big\} && \text{(interior equilibrium (mixed strategy) } x^* \in (0, 1)).
\end{aligned}$$

In the mixed region $\mathcal{R}^m$, the deceptive type chooses an interior lying frequency x^* determined by the first-order condition in Proposition 3, equivalently by the posterior y^* solving

$$\phi(y^*) = V(0) + \frac{f(\mu; \alpha)}{\delta\theta}, \quad x^* = \frac{1 - \frac{\mu}{y^*}}{1 - \mu}.$$

4.3 Graphical representation

Figure 1 displays the two threshold curves $\alpha_0(\mu)$ and $\alpha_1(\mu)$ for a particular choice of parameters $(\delta, \theta, V, \tilde{f})$. The horizontal axis is the belief μ that China is honest, and the vertical axis is the scale α of the lying payoff. The three equilibrium regions are directly visible: below $\alpha_0(\mu)$ an honesty equilibrium, between $\alpha_0(\mu)$ and $\alpha_1(\mu)$ a mixed equilibrium, and above $\alpha_1(\mu)$ a full-deception equilibrium.

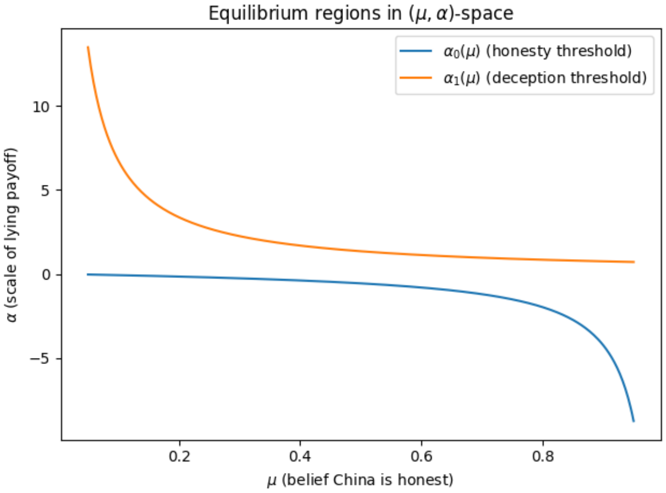

Figure 1: Threshold curves $\alpha_0(\mu)$ and $\alpha_1(\mu)$ and the three equilibrium regions.

To highlight the behavior of the deceptive type within each region, it is useful to look at the payoff function $x \mapsto U(x; \mu)$ for a fixed belief μ. Figures 2–4 plot $U(x; \mu_t)$ for three different values of α:

- a *small* lying payoff (Figure 2), where the function is decreasing and the unique maximizer is at $x^* \approx 0$ (no lying);
- an *intermediate* lying payoff (Figure 3), where $U(x; \mu_t)$ is strictly concave with a unique interior maximizer $x^* \in (0, 1)$ (mixed strategy);
- a *large* lying payoff (Figure 4), where the function is increasing and the unique maximizer is at $x^* \approx 1$ (always lying).

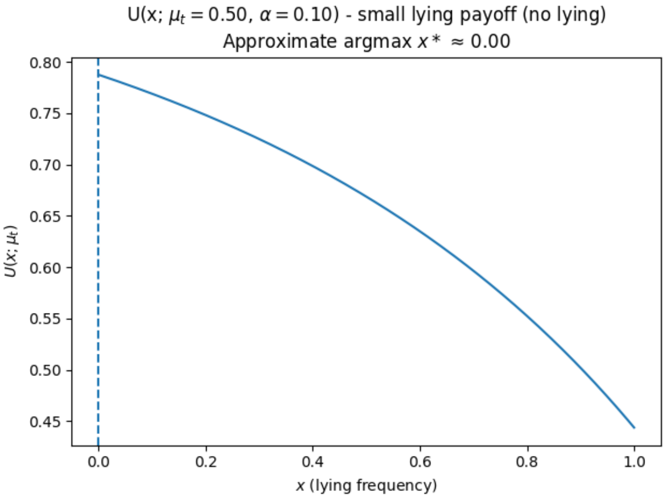

Figure 2: Deceptive type's payoff $U(x; \mu_t)$ for a small lying payoff. The unique maximizer is at $x^* \approx 0$ (honesty region).

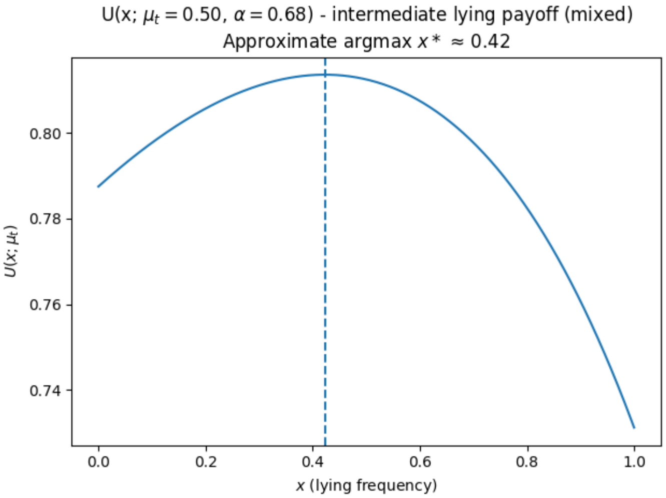

Figure 3: Deceptive type's payoff $U(x; \mu_t)$ for an intermediate lying payoff. The unique maximizer is at an interior $x^* \in (0, 1)$ (mixed region).

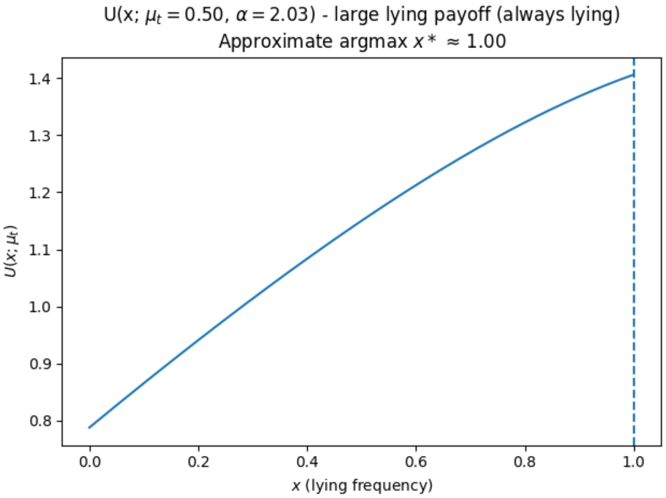

Figure 4: Deceptive type's payoff $U(x;\mu_t)$ for a large lying payoff. The unique maximizer is at $x^* \approx 1$ (full-deception region).

These figures make the algebraic equilibrium conditions transparent: in the honesty region, the derivative at zero satisfies $U'(0;\mu_t) \leq 0$ and the payoff is maximized at $x^* = 0$; in the mixed region, $U'(0;\mu_t) > 0$ and $U'(1;\mu_t) < 0$ so there is a unique interior maximizer; and in the deception region, $U'(1;\mu_t) \geq 0$ and the payoff is maximized at $x^* = 1$.

4.4 Alternative specifications for f and V

The same approach can be repeated with alternative functional forms:

- **Convex lying payoff.** If f is convex in μ (e.g. $f(\mu;\alpha) = \alpha\mu^2$), the temptation to lie grows disproportionately at high reputation levels; the mixed region $\mathcal{R}^m$ tends to expand for large μ.
- **Plateaued reputation value.** If V has a plateau near $\mu = 1$ (weak marginal returns to additional honesty at high reputation), then $V'(\mu)$ is small when μ is large. This compresses the honesty region at high μ and makes high-reputation deception more likely.
- **Steep reputation stakes.** If $V(1) - V(0)$ is large or V' is steep, the gap between $\alpha_0(\mu)$ and $\alpha_1(\mu)$ shrinks and the deception region $\mathcal{R}^1$ collapses: reputation becomes so valuable that even large immediate gains $f(\mu;\alpha)$ cannot justify systematic lying.

Thus, by choosing different parametric forms for f and V and plotting the implied curves $\alpha_0(\mu)$ and $\alpha_1(\mu)$ together with the payoff profiles $U(x;\mu_0)$, one can visually explore how the structure of reputational incentives shapes whether the equilibrium involves no lying, mixed behavior, or systematic deception.

5 Practical Insights

The comparative statics yield several implications for how a receiver such as the United States might approach monitoring and information design. Although the US's utility is not modeled explicitly, we can infer implications for US strategy by examining how the deceptive type's behavior responds to policy parameters.

First, consider the effect of changing the monitoring intensity θ. Increasing θ reduces the deceptive type's equilibrium lying frequency, since Proposition 4 established that $\frac{dx^*}{d\theta} < 0$. This reduces the frequency of misleading claims, which benefits the receiver. At the same time, the probability that a lie is detected in any period is

$$\Pr(\text{caught lie}) = \theta(1 - \mu)x^*,$$

and this expression reflects a strategic adjustment. A larger value of θ raises the chance that any given lie is checked, yet it also induces the deceptive type to lie less often. As a result, the overall detection rate may rise only modestly with θ, and very intensive monitoring can reduce the flow of informative mistakes that would otherwise expose a deceptive type. Thus the United States faces a tradeoff: deterring deception is beneficial, but eliminating opportunities for the sender to reveal itself may not be optimal. The current model does not support the US's manipulation of θ, instead treating monitoring frequency as exogenous; however, the dynamics with changes in θ are still illustrative if the US did have greater control over monitoring.

Second, it is instructive to consider whether the United States might ever benefit from a higher short-run payoff from deception. An increase in $f(\mu)$ strengthens the temptation to lie, especially for intermediate values of μ, and may induce a deceptive type to take actions that increase the likelihood of eventual detection. Although such a policy would expose the receiver to more frequent misleading claims in the short-run, it could facilitate identification of the sender's type. The desirability of such a strategy would depend on how the United States values near-term accuracy relative to the longer-term benefit of detecting a deceptive adversary.

Third, Proposition 4 indicates that the relationship between reputation and the deceptive type's lying frequency may be non-monotonic. When μ is low, the immediate benefit from deception is small, and x^* may be limited. When μ is very high, the value of preserving reputation dominates, and x^* may again be small. For intermediate values of μ, however, the deceptive type may lie more frequently. If this hump-shaped pattern obtains, the United States may prefer the sender to hold a moderate or moderately high reputation, since a deceptive sender is most likely to reveal itself in that range.

Taken together, these observations show that policies aimed at reducing deception such as increasing monitoring intensity, lowering the short-run gains from lying, or reducing the sender's reputation are not always unambiguously beneficial. The deceptive type adjusts its behavior, and these adjustments can reduce the likelihood that a deceptive sender ever reveals itself. In some environments, the receiver may find it optimal to employ intermediate monitoring, to tolerate or even enhance the sender's reputation, or to allow higher short-run incentives for deception, in order to increase the probability of eventual detection.

6 Limitations and future research

The model highlights the central strategic tradeoff but relies on several simplifying assumptions. One important simplification is that a single detected lie permanently reduces reputation to zero. This assumption makes the incentives transparent but likely overstates the permanence of reputational damage. In many real settings credibility can recover gradually, and an extension that allows partial or slow reputation repair would provide a richer account of long-run dynamics. Additionally, not all lies are equal. Introducing a 'magnitude' parameter for deception (small misrepresentations vs. big lies) would make the model more realistic. Real-world behavior includes hedged falsehoods and slight misrepresentations, not just binary truths or lies.

A second simplification is the use of a reduced-form continuation value $V(\mu)$ rather than a continuation value derived from an explicit infinite-horizon game. A fully microfounded dynamic structure, in which the belief μ evolves as part of a fixed-point distribution of future play, could yield additional insight into how reputation accumulates and erodes over time.

Furthermore, in the current setup the receiver is passive: we do not model any strategic choice or payoff for the US. The Practical Insights section extrapolates potential strategic directions given the model, but the model itself does not formalize these. Endogenizing the receiver's strategy (e.g., making θ a choice variable or allowing costly punishment) is a promising direction for future research.

Finally, the model focuses on a single sender and a single receiver. Real-world situations often involve multiple audiences, including domestic actors, third-party observers, and foreign governments whose beliefs may evolve in different ways. Extending the model to incorporate multiple receivers or cross-audience reputational spillovers would offer an important avenue for future work.

Despite these simplifications, the model provides a tractable environment in which the main reputational incentives can be clearly understood. Future research could extend the framework to include partial reputation recovery or gradations of deception severity. Another avenue would be to apply the model's predictions to historical case studies, to test how well the theory explains empirical patterns.

7 Empirical Relevance and Observable Implications

The model is motivated by China's signaling over Taiwan and in the South China Sea, but it also refines and extends the lessons that policymakers and analysts have drawn from these episodes. In the motivating examples, China mixes truthful disclosures with misleading claims, and verification is infrequent (Gunness & Saunders, 2022; Tiong & Hoo, 2024). Our model formalizes this pattern, showing it to be the result of an optimal tradeoff rather than ad-hoc behavior. An authoritarian sender with informational control optimally chooses a lying frequency that balances short-run gains from deception against the risk of a discrete reputational collapse once a lie is verifiably exposed. This yields new predictions about when deception should spike, when it should recede, and how changes in monitoring and patience shift that balance.

The Spratly Islands case illustrates this tradeoff. In 2015, President Xi Jinping publicly pledged that China had "no intention to pursue militarization" of the artificial islands, a statement that was initially treated as credible but later contradicted by commercial satellite imagery showing runways and air defenses on the outposts (Panda, 2016). Subsequent commentary described a sharp erosion of Beijing's credibility on maritime issues (Stokes, 2019). In terms of the model, this is an instance where a deceptive type (China) selected a high lying frequency while effective monitoring θ was perceived to be low, then suffered a reputational collapse once satellite and open-source intelligence raised the true detection probability. As a result, later assurances about Taiwan and regional posturing are heavily discounted as μ_{t+1} is driven to zero.

The model also helps interpret the more nuanced pattern of partial truth-telling and calibrated ambiguity that followed improvements in verification. Analyses of Chinese statements and behavior around the Spratlys document how rhetoric shifted from categorical denials of militarization toward phrasing that emphasized dual-use or defensive functions once satellite imagery became widely available (Panda, 2016; Asia Maritime Transparency Initiative, as discussed in Stokes, 2019). In the theory, this corresponds to an interior equilibrium lying frequency x^* in which the deceptive type mixes between truthful and misleading claims. The sender exploits ambiguity but avoids fully exhausting its remaining stock of credibility. This goes beyond the simple observation that authoritarians sometimes lie and generates the more specific prediction that mixed-strategy deceptive signaling is most likely when reputation is intermediate and monitoring is imperfect but nontrivial.

The comparative statics on the monitoring parameter θ map directly onto an empirical literature on verification technologies. The diffusion of commercial satellite imagery and other open-source intelligence has dramatically increased the ability of governments, journalists, and civil society to monitor military deployments, treaty compliance, and environmental performance (Jo, 2019; Marleku, 2025). Work using satellite night-light data shows that authoritarian regimes systematically overstate reported GDP growth relative to independent proxies, and that improved external monitoring allows researchers to recover these discrepancies (Martínez, 2022). In the language of the model, such technologies raise effective θ, which the theory predicts will reduce the equilibrium frequency of brazen, easily falsifiable lies and shift behavior toward more carefully framed, narrower claims. This is consistent with the observed evolution from unconditional pledges to statements that are technically defensible yet strategically misleading once monitoring improves.

The same logic appears in other domains where deception is periodically audited. In anti-doping, stronger monitoring regimes that retain biological samples for years and retest them with improved techniques are associated with higher eventual detection rates and lower simulated prevalence of doping, as athletes anticipate a higher probability that violations will eventually be uncovered (Westmattelmann et al., 2025). In corporate finance, the tightening of audit standards and internal controls after the Sarbanes–Oxley Act reduced overt accrual-based earnings manipulation and shifted firms toward subtler real earnings management (Cohen, Dey, & Lys, 2008). These cases match the model's prediction that increases in θ reduce blatant, easy-to-detect misrepresentation but may redirect deceptive effort into less verifiable margins rather than eliminating it altogether. This could be further developed in models that give greater nuance to lying severity instead of treating it as a binary outcome.

Technological change does not unambiguously increase θ, however. The same advances in machine learning and digital media that enable better verification also lower the cost of producing sophisticated false content. Intelligence and policy analyses describe deepfakes and AI-assisted disinformation as an "autocrat's new toolkit" that can temporarily overwhelm verification capacities and exploit lags in institutional adaptation (Canadian Security Intelligence Service, 2023; Chesney & Citron, 2019; Fontaine & Frederick, 2019). In model terms, these innovations can reduce effective monitoring, θ, if detection tools and norms do not keep pace. The empirical literature on information operations around the Ukraine war shows both sides of this dynamic: public intelligence disclosures and open-source sleuthing have been used to preempt and counter false narratives, while adversaries simultaneously experiment with new techniques that stress existing verification systems (Marleku, 2025). Therefore, the net effect of technology on θ is ambiguous ex ante and that the direction of change is an empirical matter.

Taken together, these examples show that the model both formalizes intuitions that practitioners already use and generates new, testable predictions. First, deception should be most intense when monitoring is imperfect but nonnegligible and when the sender's reputation is neither fully destroyed nor pristine. Second, sustained improvements in verification capacity, whether through satellite imagery, audits, or long-horizon testing, should reduce overt lying and redirect deceptive effort into narrower, more ambiguous practices. Third, technological shocks that raise deceptive capacity faster than verification capacity should temporarily increase effective lying payoffs and produce spikes in manipulation, particularly in the dimensions that remain hardest to check. These predictions can be evaluated using historical episodes in the Taiwan Strait and South China Sea, cross-country evidence on authoritarian data manipulation, and sector-specific studies of doping and financial reporting, linking the theory to a growing empirical literature on information control and reputational constraint.

8 Conclusion

The model links China's signaling over Taiwan and the Spratlys to a simple reputational logic: an authoritarian sender with asymmetric control over information chooses a lying frequency that trades off immediate gains from manipulation against the risk of a catastrophic loss of credibility when a lie is exposed. In doing so, the model provides a tractable equilibrium characterization of strategic ambiguity, pinpointing how often a sender will lie in equilibrium as a function of monitoring, patience, and payoffs. Monitoring intensity and patience push toward honesty by raising the shadow value of reputation, while stronger short-run returns to deception push in the opposite direction. When these forces balance in the intermediate region, the deceptive type mixes, exploiting ambiguity without fully exhausting its stock of credibility.

Although the empirical motivation comes from US-China rivalry, the underlying mechanism is more general. Any setting that combines repeated signaling, unequal access to the information environment, and occasional verifiable checks will exhibit similar reputational dynamics: an energy exporter managing information about production shocks, a dominant technology platform communicating about content moderation practices to regulators, or an authoritarian government signaling compliance with human-rights or arms-control commitments to international monitors. In all of these cases, the framework highlights how the design of monitoring institutions and verification technologies (who can credibly check claims, how often, and at what cost) feeds back into the equilibrium use of deception and the long-run distribution of influence.

9 References

- Al Jazeera. (2023, September 29). US says China's 'global information manipulation' threatens freedoms. *Al Jazeera.*

- Canadian Security Intelligence Service. (2023, May 24). Finding signals in the synthetic: Intelligence in the era of deepfakes. In *The evolution of disinformation: A deepfake future* (pp. 57–68).

- Chesney, R., & Citron, D. K. (2019). Deep fakes: A looming challenge for privacy, democracy, and national security. *California Law Review*, 107(6), 1753–1820.

- Cohen, D. A., Dey, A., & Lys, T. Z. (2008). Real and accrual-based earnings management in the pre- and post-Sarbanes–Oxley periods. *The Accounting Review*, 83(3), 757–787.

- Crawford, V. P., & Sobel, J. (1982). Strategic information transmission. *Econometrica*, 50(6), 1431–1451.

- Fontaine, R., & Frederick, K. (2019, March 15). The autocrat's new tool kit. *The Wall Street Journal.*

- Gunness, K., & Saunders, P. C. (2022, December 22). *Averting escalation and avoiding war: Lessons from the 1995–1996 Taiwan Strait crisis* (China Strategic Perspectives 17). Washington, DC: NDU Press.

- Jo, S. (2019, January 30). Satellites reveal the truth: China has failed to reduce greenhouse gas emissions. *DongA Science*.

- Marleku, A. (2025). Public intelligence as a strategic tool: The role of real-time intelligence disclosure in the Ukraine War. *Security and Defence Quarterly*, 50(2), 50–65.

- Martínez, L. R. (2022). How much should we trust the dictator's GDP growth estimates? *Journal of Political Economy*, 130(10), 2731–2769.

- Panda, A. (2016, December 16). It's official: Xi Jinping breaks his non-militarization pledge in the Spratlys. *The Diplomat*.

- Stokes, J. (2019, August 6). China's credibility problem. *Defense One*.

- Tiong, W. J., & Hoo, T. B. (2024, April 21). China's strategic ambiguity on Taiwan. *SAIS Review of International Affairs*.

- Westmattelmann, D., Sprenger, M., Lanfer, J., Stoffers, B., & Petróczi, A. (2025). The impact of sample retention and further analysis on doping behavior and detection: Evidence from agent-based simulations. *Frontiers in Sports and Active Living*, 7, 1578929.

A Code for equilibrium illustrations

```
import numpy as np
import matplotlib.pyplot as plt

# ------------------------------
# 1. Parameters and primitives
# ------------------------------
delta = 0.9
theta = 0.5
v0 = 0.0
v1 = 2.0
gamma = 1.0

def V(mu):
    """Reputation value function: quadratic, increasing, concave."""
    return v0 + v1 * mu - 0.5 * gamma * mu**2

def V_prime(mu):
    return v1 - gamma * mu
```

```
def f_tilde(mu):
    """Baseline lying payoff shape; we scale it by alpha."""
    return mu  # simple increasing function

# ------------------------------
# 2. Threshold curves alpha0(mu), alpha1(mu)
# ------------------------------
V0 = V(0.0)
V1 = V(1.0)

def alpha0_func(mu):
    # Honesty threshold from the condition for x* = 0
    return delta * theta * (V(mu) - V0 - V_prime(mu) * (mu / (1 - mu))) / f_tilde(mu)

def alpha1_func(mu):
    # Deception threshold from the condition for x* = 1
    return delta * theta * (V1 - V0) / f_tilde(mu)

mu_grid = np.linspace(0.05, 0.95, 400)
alpha0 = alpha0_func(mu_grid)
alpha1 = alpha1_func(mu_grid)

plt.figure()
plt.plot(mu_grid, alpha0, label=r’$\alpha_0(\mu)$ (honesty threshold)’)
plt.plot(mu_grid, alpha1, label=r’$\alpha_1(\mu)$ (deception threshold)’)
plt.xlabel(r’$\mu$ (belief China is honest)’)
plt.ylabel(r’$\alpha$ (scale of lying payoff)’)
plt.title(r"Equilibrium regions in $(\mu, \alpha)$-space")
plt.legend()
plt.tight_layout()
plt.show()

# ------------------------------
# 3. U(x; mu, alpha) for a fixed mu0
# ------------------------------
def mu_update(mu, x):
    """Bayesian update after truthful claim, given mu and x."""
    return mu / (1 - x + x * mu)  # = mu / (1 - x(1-mu))

def U_of_x(x, mu, alpha):
    """Deceptive type’s payoff U(x; mu, alpha)."""
    mu = float(mu)
    V_mu = V(mu)
    V0_local = V(0.0)
    f_val = alpha * f_tilde(mu)
    # Payoff from lying in this period
    L = f_val + delta * (1 - theta) * V_mu + delta * theta * V0_local

    # Payoff from honesty in this period (given that long-run lying freq is x)
    mu_next = mu_update(mu, x)
    H = delta * theta * V(mu_next) + delta * (1 - theta) * V_mu

    return x * L + (1 - x) * H

# Choose a representative mu0
mu0 = 0.5

alpha0_mu0 = alpha0_func(mu0)
alpha1_mu0 = alpha1_func(mu0)

print("At mu = 0.5:")
print("alpha0(mu) =", alpha0_mu0)
print("alpha1(mu) =", alpha1_mu0)

# Illustrative cases
alpha_low = 0.1                 # small lying payoff -> x* ~ 0
alpha_mid = alpha1_mu0 / 2.0    # intermediate -> mixed
alpha_high = alpha1_mu0 * 1.5   # large -> x* ~ 1

x_grid = np.linspace(0.0, 1.0, 400)

def plot_U_for_alpha(alpha, title_suffix):
    U_vals = np.array([U_of_x(x, mu0, alpha) for x in x_grid])
    x_star = x_grid[np.argmax(U_vals)]
    plt.figure()
    plt.plot(x_grid, U_vals)
    plt.axvline(x_star, linestyle=’--’)
    plt.xlabel(r’$x$ (lying frequency)’)
    plt.ylabel(r’$U(x;\mu_t)$’)
    plt.title(
        rf’U(x; $\mu_t={mu0:.2f}$, $\alpha={alpha:.2f}$) - {title_suffix}’ + ’\n’ +
        rf’Approximate argmax $x*$  {x_star:.2f}’
    )
    plt.tight_layout()
    plt.show()

plot_U_for_alpha(alpha_low, "small lying payoff (no lying)")
plot_U_for_alpha(alpha_mid, "intermediate lying payoff (mixed)")
plot_U_for_alpha(alpha_high, "large lying payoff (always lying)")
```