

Refusal Externalities: Non-Evasive Refusals as an Information Public Good in Dynamic AI Safety

Liz Lemma

Future Detective

January 22, 2026

Abstract

Modern alignment systems (e.g., Constitutional AI’s non-evasive harmless assistant and Safe RLHF’s constrained optimization) highlight a practical tension: evasive refusals reduce immediate harm but also reduce transparency and the ability to discover and patch vulnerabilities through red teaming. We formalize this as a dynamic economic problem in which an AI developer chooses a refusal style for risky prompts—either evasive refusal or explanatory refusal that engages safely and clarifies why the request is harmful. Explanatory refusals generate monitoring-relevant information (higher vulnerability discovery rates), accelerating future robustness improvements; evasive refusals slow learning and can increase long-run harm. In a tractable Markov model with affine learning dynamics and linear payoffs, we show the developer’s equilibrium refusal policy is characterized by a threshold: explain early when the system is fragile (few vulnerabilities discovered) and switch to evasiveness only when robustness is sufficiently high. We further show a Refusal Externality: if the informational benefits of explanatory refusals spill over to auditors, regulators, and the broader ecosystem, private developers underprovide explanation relative to the social optimum. The model yields comparative statics in audit intensity, attacker innovation, harm severity, and evaluation penalties for evasiveness—providing economic justification for 2026-era standards and procurement requirements that prefer non-evasive, explanatory refusals when safe. We outline an empirical design comparing RLHF regimes that reward explanation vs evasiveness, measuring vulnerability discovery rates, adaptive red-team robustness, and long-run harm.

Table of Contents

1. Introduction: why refusal style is now an economic design variable (safety ops, liability, trust) and how CAI/Safe RLHF motivate a dynamic perspective beyond static harmlessness metrics.

2. 2. Institutional background: refusal/evasiveness in RLHF and CAI; why evaluation instructions matter; the monitoring/red-teaming pipeline as an information production process.
3. 3. Model setup: state of discovered vulnerabilities; refusal actions; learning/obsolescence dynamics; per-period benefits and harms; optional informational spillover term capturing ecosystem value.
4. 4. Equilibrium concept: developer's Markov decision problem; Bellman equation; conditions for monotone (threshold) optimal policy; existence and uniqueness discussion.
5. 5. Main results: Refusal Externality Theorem; closed-form threshold under one-switch/linear-value approximations; welfare comparison (private vs social planner).
6. 6. Comparative statics: audit intensity, attacker innovation (obsolescence), harm severity, and evaluation penalties for evasiveness; when standards should mandate explanation.
7. 7. Extensions (optional, tractable): (i) endogenous attack frequency, (ii) heterogeneous users (benign vs malicious), (iii) multi-category harms with different learning rates, (iv) partial explanations vs full explanations.
8. 8. Empirical design sketch: operationalizing 'discovery rate' and 'adaptive robustness'; training regimes (reward explanatory refusals vs evasive refusals); outcome metrics and identification strategy.
9. 9. Discussion and policy implications: procurement language, audit design, transparency requirements; limits of the model and what would require numerical methods.
10. 10. Conclusion.

1 Introduction: refusal style as an economic design variable

Refusals on risky prompts used to be treated as an essentially “static” safety feature: a model either does or does not comply with requests for wrongdoing, and evaluation focuses on the immediate content of a single response. In current deployments, we increasingly find that *how* a model refuses—whether it offers a brief, evasive non-answer or a more explanatory, policy-grounded refusal—has become a first-order design choice with operational, legal, and reputational consequences. We can no longer treat refusal behavior as a thin wrapper around a fixed capability; it is a control surface that shapes user behavior, incident response, and the system’s own improvement loop.

Three forces make refusal style economically salient.

First, refusal style is now part of *safety operations*. A deployment organization typically runs a monitoring and response pipeline: telemetry on flagged conversations, incident triage, red-teaming and adversarial testing, patching (via prompt updates, data curation, system-level classifiers, or fine-tuning), and re-evaluation. Refusal outputs are not merely outcomes to be scored; they are data products that enter this pipeline. An explanatory refusal may reveal which safety boundary was triggered (e.g., chemical synthesis assistance, evasion of safeguards, or targeted persuasion) and may articulate the model’s “understanding” of the request in a structured way that is useful for auditors and engineers. An evasive refusal, by contrast, may minimize immediate leakage but also collapses information about failure modes into a generic “I can’t help with that.” If we care about reducing future harm, we should treat refusals as part of a feedback system that allocates attention and accelerates (or slows) vulnerability discovery.

Second, refusal style affects *liability and governance exposure*. Regulators, courts, and procurement processes increasingly evaluate not only whether a model avoids facilitating wrongdoing, but also whether it behaves in a way that is transparent, consistent, and auditable. “Policy citing” refusals, which explicitly reference constraints and offer safe alternatives, can demonstrate due care and reduce ambiguity about the system’s intended use. At the same time, detailed refusals can be framed as negligent if they provide actionable scaffolding for misuse (e.g., by narrowing the search space for an attacker). This creates a genuine economic tradeoff: the refusal policy influences both the probability and severity of harmful incidents *and* the expected penalty conditional on incidents, through foreseeability, documentation, and the organization’s ability to show effective controls.

Third, refusal style shapes *trust and user utility* in benign contexts. In many products, the marginal user cost of an overly evasive refusal is not just frustration; it is reduced reliance, lower retention, and migration to less safe alternatives. Explanatory refusals can preserve helpfulness by redirecting the

user toward safe substitutes (e.g., high-level safety information, compliance-friendly guidance, or harm-minimizing resources). They can also reduce the risk that benign users repeatedly probe the boundary out of confusion, which itself generates noisy monitoring load. Thus, refusal style mediates a familiar platform design tension: safety interventions that are too blunt impose user-experience costs that may be privately salient to the developer even when the direct harm avoided is socially salient.

These considerations motivate a dynamic perspective that is not captured by static harmlessness metrics. Many standard benchmarks effectively ask: “Given a risky prompt, does the model output disallowed content?” This is necessary but insufficient for two reasons. First, harm is not a one-shot event: attackers iterate, defenders patch, and the underlying threat model evolves. Second, the content of refusals changes the *rate* at which both attackers and defenders learn. A refusal policy that minimizes immediate leakage could still be dynamically suboptimal if it slows the system’s ability to discover and mitigate vulnerabilities faster than the adversary can innovate. Conversely, a refusal policy that explains too much could be locally attractive (e.g., for user satisfaction) while increasing near-term misuse risk. We therefore need a formalism that can express the intertemporal tradeoff between (i) immediate exposure and (ii) future robustness through information production.

Constitutional AI (CAI) and “Safe RLHF”-style methods make this trade-off concrete. CAI encourages the model to follow a written constitution and to justify its behavior in ways that are legible to users and auditors. In practice, this often produces refusals that are more explicit about principles (e.g., harm prevention, privacy, illegality) and that offer safer alternatives. Safe RLHF pipelines similarly train models to avoid unsafe content while preserving helpfulness, frequently resulting in refusals that include explanations, boundary statements, and redirections. These approaches are attractive because they can make safety behavior more consistent and easier to evaluate, but they also heighten the question of whether the refusal output itself becomes a channel for capability transfer. Put bluntly: a refusal that “explains” why something is harmful can sometimes be indistinguishable from a high-level recipe for how to do it.

The key observation is that refusal style changes the *informativeness* of interactions for multiple stakeholders. Users receive information (which may deter misuse, or may help them rephrase attacks). The developer receives information via logs, incident reports, and model behavior under probing. External auditors and the broader ecosystem may also learn, through shared evaluations, bug reports, and norms that propagate across organizations. When explanatory refusals make it easier to localize failures, they can increase the rate at which vulnerabilities are discovered and mitigated. But those benefits need not be fully internalized: ecosystem-level learning, reduced systemic risk, and shared best practices are partially public goods. This suggests a potential wedge between privately optimal refusal style and

socially optimal refusal style, even holding fixed the developer’s direct harm costs.

A dynamic framing also makes room for an often-ignored feature of real deployments: *obsolescence*. Threat models drift, jailbreak techniques mutate, and newly connected tools expand the action space. In a rapidly changing environment, robustness is not a monotone achievement; it is an asset that depreciates unless continuously replenished by discovery and mitigation. This is precisely the setting in which the informational content of refusals can matter: if we expect continuous attacker innovation, then policies that speed defensive learning can have disproportionately large long-run value.

At the same time, we should not romanticize explanation. There are clear failure modes. Explanatory refusals can be gamed: attackers can treat them as an oracle for boundary-finding, extracting a taxonomy of constraints and then searching for paraphrases that evade them. Explanations can leak model-internal heuristics (e.g., what content triggers filters), enabling targeted bypass. They can also lead to “capability laundering” if safe alternatives are too close to the disallowed goal. Finally, explanation may create governance risks when it appears to provide advice, even if intended as deterrence. Any useful formal model must therefore allow an *immediate* harm term from explanation that can be larger than that of evasiveness.

Our goal in the remainder of this work is to make these tradeoffs explicit in a minimal dynamic control problem. We treat refusal style as a policy lever that affects both (i) contemporaneous harm and (ii) the evolution of system robustness through discovery and mitigation. This lens does not assume that explanatory refusals are always good or that evasive refusals are always bad; rather, it clarifies when each is optimal, how thresholds depend on measurable parameters (learning rates, depreciation, and penalty intensity), and where externalities arise.

Before introducing the formal model, we next ground the discussion in institutional practice: how RLHF and CAI training pipelines operationalize refusal and non-evasiveness, why evaluation instructions (and procurement criteria) can implicitly price refusal style, and how the monitoring/red-teaming pipeline functions as an information production process that interacts with refusal outputs.

2 Institutional background: refusal and evasiveness as training and evaluation targets

In practice, refusal style is not chosen in a vacuum at deployment time; it is the end product of training objectives, policy documents, and evaluation regimes that implicitly (and sometimes explicitly) assign value to being explanatory versus being terse. We find it useful to treat these institutional choices as part of the “mechanism” that prices different refusal behav-

iors. This section summarizes how contemporary RLHF/“helpful-harmless” pipelines and Constitutional AI (CAI) operationalize refusal, why evaluator instructions can change the learned equilibrium, and how monitoring and red-teaming can be understood as an information production process whose effectiveness depends on the content of refusals.

A typical RLHF stack begins with supervised fine-tuning (SFT) on demonstrations, followed by preference modeling and reinforcement learning (or related direct optimization) against a learned reward signal ????. Refusal behavior enters at all three stages. During SFT, demonstrations often include templated refusals: a short statement of inability, a brief policy rationale (e.g., illegality, harm, privacy), and a redirect to permissible alternatives. During preference data collection, labelers are commonly instructed to down-rank “stonewalling” responses that are unhelpful even when they are safe, and to up-rank refusals that are polite, firm, and provide safe substitutions. When these preferences are distilled into a reward model, they create a measurable incentive to avoid purely evasive non-answers. In other words, non-evasiveness can become a trained capability: models learn to produce refusals that are more structured, more consistent, and more legible to humans.

However, the same machinery can also create incentives for over-disclosure. When labelers reward specificity (because specificity correlates with perceived helpfulness on benign queries), the model can learn to be specific even in refusal contexts—for example by naming the prohibited category, clarifying what counts as “actionable,” or enumerating disallowed sub-steps before declining. From a safety standpoint, such content can be dual-use: it may deter benign users by clarifying boundaries, yet it can also function as a boundary-finding oracle for attackers. This is not merely a theoretical concern. In operational red-teaming, we routinely observe iterative attack strategies that treat refusal text as an informative channel: the attacker probes, reads the refusal rationale, and adapts the next prompt to target the apparent decision rule (e.g., shifting from “instructions” to “fiction,” or from “how to” to “what are common mistakes”). Thus, the reward shaping that produces “better” refusals under human preferences may also increase the informational gradient available to adversaries.

CAI-style training makes these incentives even more explicit by encouraging models to justify their behavior with reference to a written constitution ?. The CAI workflow typically uses self-critique and revision: the model produces an initial response, critiques it against principles, and then outputs a revised response. In deployments influenced by CAI, refusals often contain (i) a principle citation (harm prevention, illegality, privacy, non-violent wrongdoing), (ii) an explanation of why the request violates the principle, and (iii) a safer alternative. This can be attractive for governance: principle-grounded refusals are easier to audit, can be made more consistent across languages and domains, and can be aligned with external policy commitments (e.g.,

“we do not provide instructions for cyber intrusion”). But CAI also raises a design question: when a model is trained to be explicitly reason-giving, it may reveal the structure of the safety policy in ways that facilitate evasion. Even when the refusal contains no explicit procedural instructions, it may narrow the adversary’s search over attack prompts by indicating what the system “noticed” (targets, intent, quantities, or contextual cues). Put differently, CAI can increase the mutual information between the attacker’s probe and the defender’s observable output, which is desirable for internal debugging but potentially harmful under adversarial adaptation.

Evaluation instructions and procurement criteria play a second, underappreciated role: they determine what training teams optimize for and what product teams are rewarded for shipping. Many benchmark protocols include an explicit non-evasiveness axis—e.g., penalizing refusals that are generic, inconsistent, or unhelpful, and rewarding refusals that provide safe alternatives and clear boundaries. This matters because RLHF pipelines are often “eval-driven”: teams iterate on prompts, policies, and training data to move the evaluation score. If an evaluation suite rewards explanatory refusals (for being policy-consistent and user-friendly), the platform acquires a private benefit to explanation independent of any safety dynamics. Conversely, if evaluation or legal review penalizes any content that could be construed as facilitating wrongdoing, teams may adopt conservative guidance that favors terse refusals, even at the cost of user frustration and reduced auditability. These incentives are not hypothetical; they show up in vendor questionnaires, model risk management templates, and regulator-facing documentation, where organizations must demonstrate both “robustness to misuse” and “transparency of controls.” The key point is that non-evasiveness can be priced either positively (as a trust and usability feature) or negatively (as a perceived liability), and those prices feed back into the learned refusal policy.

We can also view evaluation instructions as shaping what signals are available to the organization’s own monitoring pipeline. Consider two extreme labeler guidelines. One says: “Refuse briefly; do not mention the specific policy category; avoid any detail that could guide misuse.” The other says: “Refuse with a clear policy reason and provide a safe alternative; be consistent and explicit about boundaries.” Both can yield high safety on a narrow “no disallowed content” metric, but the second produces richer structured text that can be mined for debugging: it often contains a self-classification of why the model refused, which can be used to route incidents, cluster failure modes, and measure drift. In organizations with large volumes of user interactions, this difference can materially change the marginal cost of triage and the speed with which engineers identify systematic gaps (e.g., a new jailbreak pattern that the refusal itself describes as “roleplay” or “educational” framing). Thus, evaluator instructions influence not only the outward user experience but also the internal observability of safety-relevant events.

Finally, the monitoring and red-teaming ecosystem is itself an information production process, and refusal style affects its productivity. In mature deployments, the operational loop typically includes (i) automated detection (classifiers, keyword triggers, anomaly detection), (ii) logging and sampling policies, (iii) human review and incident escalation, (iv) targeted red-teaming to reproduce and generalize failures, (v) mitigation (policy updates, classifier changes, data collection, fine-tuning), and (vi) regression evaluation. External auditors, bug bounty programs, and shared evaluation sets can be understood as additional sources of discovery that sometimes spill across organizational boundaries. Within this loop, a refusal is not merely a “safe outcome”; it is a data point that can either preserve or destroy information about where the model sits relative to safety boundaries. Explanatory refusals can increase the yield of this pipeline by making it easier to infer latent intent categories, to detect near-misses (cases where the model almost complied), and to measure distribution shift (new topics or new attack framings that produce qualitatively different refusal rationales). Evasive refusals, by design, tend to compress these signals, which can reduce the immediate risk of leakage but can also slow down learning from incidents and near-incidents.

This framing suggests a useful abstraction: refusal behavior controls how effectively risky interactions are converted into actionable knowledge for defenders, even holding fixed the underlying model capability. At the same time, we must acknowledge the failure mode that motivates evasiveness: the same information that helps defenders may also help attackers, and the net effect is ambiguous *ex ante*. The operational choice is therefore not “explain or do not explain” in the abstract, but rather how to allocate explanatory bandwidth across time and states of system robustness, subject to an evolving threat model and imperfect incentives. In the next section, we formalize this intuition with a minimal dynamic model in which refusal style affects both contemporaneous harm and the rate of vulnerability discovery and mitigation, allowing us to derive threshold policies and to characterize when private incentives diverge from social objectives.

3 Model setup: state, refusal actions, learning dynamics, and payoffs

We now formalize the safety–usability tradeoff implicit in refusal style. The core idea is that what the model says when it refuses is not merely a present-tense interaction outcome; it is also an input into an ongoing discovery-and-mitigation pipeline. Explanatory refusals can create structured signals that make it easier to diagnose boundary failures, cluster incidents, and harden the system, but they can also leak information that increases short-run risk. Evasive refusals reduce what is revealed in the moment, but may slow down the rate at which risky interactions are converted into actionable knowledge.

State: a reduced-form “robustness stock.” Time is discrete, indexed by $t = 0, 1, 2, \dots$. The system state is a scalar $x_t \in [0, 1]$, interpreted as the fraction of an underlying (normalized) “vulnerability mass” that has been discovered and mitigated by time t . Higher x_t means the deployed model-and-mitigation stack is more robust in the sense that fewer exploitable weaknesses remain. The residual mass $1 - x_t$ summarizes all ways the system could still fail on risky prompts (e.g., jailbreakable patterns, policy gaps, monitoring blind spots, or model behaviors that enable misuse). This aggregation is deliberate: our goal is not to track individual vulnerabilities, but to capture how refusal policy affects the *rate* at which the platform learns about and closes them.

We assume x_t is observed by the decision-maker at the start of period t . Concretely, x_t can be thought of as a sufficient statistic for internal safety indicators (incident rates, red-team findings closed, coverage of known attack families), compressed to a single dimension.

Actions: refusal style as a control variable. In each period, the platform chooses a refusal style $a_t \in \{E, R\}$ applied to the risky-prompt environment:

$$a_t = \begin{cases} E & \text{explanatory refusal (reason-giving, boundary clarification, safe alternatives),} \\ R & \text{evasive refusal (terse, generic, minimal information).} \end{cases} \quad (1)$$

We treat benign prompts as contributing an additive constant to payoffs and omit them without loss of generality. The action a_t should be read as the outcome of an institutional bundle (training targets, style guidelines, evaluator preferences, and policy constraints) that pins down how the system responds *conditional on refusing*. The modeling choice to focus on the refusal channel isolates the mechanism we care about: the information content of safety enforcement.

Learning and obsolescence dynamics. Refusal style affects the efficiency with which the platform (and its surrounding ecosystem) discovers and mitigates the remaining vulnerability mass. We encode this with an action-dependent “learning/discovery rate” $\alpha_a \in [0, 1]$. We assume

$$\alpha_E > \alpha_R \geq 0, \quad (2)$$

capturing that explanatory refusals produce richer artifacts (structured rationales, self-classifications, more legible boundary descriptions) that can be mined by monitoring pipelines and red teams, while evasive refusals compress those signals. The key is not that evasive refusals generate zero information, but that they reduce marginal learnability.

At the same time, robustness is not purely cumulative: threat actors innovate, new domains are deployed, and mitigations can decay. We incorporate this with an “obsolescence/attacker-innovation” rate $\omega \in [0, 1)$. When ω is higher, a larger fraction of accumulated robustness becomes ineffective each period (or equivalently, new vulnerability mass arrives that must be rediscovered).

We model the law of motion as the affine transition

$$x_{t+1} = f_{a_t}(x_t) := (1 - \omega)[(1 - \alpha_{a_t})x_t + \alpha_{a_t}]. \quad (3)$$

This functional form has three features we will use throughout. First, it is increasing in x_t : a more robust system remains (weakly) more robust next period, holding the refusal style fixed. Second, the state drifts toward 1 at a rate governed by α_a , but only insofar as robustness is not eroded by ω . Third, the *incremental* learning advantage of explanatory refusals shrinks as x_t rises:

$$f_E(x) - f_R(x) = (1 - \omega)(\alpha_E - \alpha_R)(1 - x), \quad (4)$$

so explanation matters most when there is “more left to learn” (low x) and becomes less pivotal as the system approaches the frontier $x \rightarrow 1$. This diminishing-returns structure is meant to capture that once major vulnerability families are patched and monitoring is mature, additional reason-giving yields less marginal discovery.

It is also helpful to note the boundary behavior. When $x = 1$, we have $f_a(1) = (1 - \omega)$, so even a fully hardened system is pulled below perfection when $\omega > 0$: continual learning is required to keep pace. When $x = 0$, we obtain $f_a(0) = (1 - \omega)\alpha_a$, so the first steps in hardening are governed directly by the learning rate induced by refusal style.

Per-period platform payoffs: benefits of style and costs of harm.

We specify a per-period private payoff that combines (i) a benefit term reflecting user trust and product utility associated with the refusal style, and (ii) a cost term proportional to expected harm from residual vulnerability and any immediate leakage induced by the refusal content. Formally, under action $a \in \{E, R\}$ at state x :

$$\pi_a(x) = b_a - \kappa(\bar{h}(1 - x) + \Delta_a). \quad (5)$$

Here b_a is an action-specific “experience” or “reputational” benefit from the refusal style. A natural interpretation is that explanatory refusals reduce user frustration, increase perceived transparency, and perform better on non-evasiveness evaluations, so one may have $b_E > b_R$, though we do not require this. The term $\bar{h}(1 - x)$ is baseline harm intensity: as long as residual vulnerability mass remains, there is some chance that risky prompts lead to harmful outcomes (successful jailbreaks, facilitated misuse, or operational

incidents). The parameter $\kappa > 0$ scales this harm into the platform’s private objective (e.g., via expected liability, enforcement risk, reputational damage, or internalized ethical cost).

Finally, $\Delta_a \geq 0$ captures *incremental immediate harm* directly attributable to the refusal style, holding x fixed. The canonical case is $\Delta_E \geq \Delta_R$: explanatory refusals may leak boundary information or provide a more informative oracle for adversarial adaptation, increasing short-run risk even if they also improve long-run learning. Importantly, by separating $\bar{h}(1 - x)$ from Δ_a , we allow immediate leakage risk to persist even in relatively robust states (e.g., some information channels remain sensitive regardless of overall hardening).

A social objective with informational spillovers. The platform typically does not capture all benefits from faster discovery. Mitigations, benchmarks, incident taxonomies, and red-team methods often spill into the broader ecosystem (through publications, shared vendor practices, open-source tools, and regulator learning), reducing systemic risk beyond the focal platform. To represent this, we introduce a social planner payoff that adds a public-good term proportional to the extra discovery induced by explanatory refusals relative to the evasive baseline:

$$\tilde{\pi}_a(x) = \pi_a(x) + g \cdot (\alpha_a - \alpha_R)(1 - x), \quad (6)$$

where $g \geq 0$ is the marginal social value of incremental discovery. The term $(\alpha_a - \alpha_R)(1 - x)$ isolates the portion of learning attributable to choosing E rather than the baseline R , and scales it by the remaining vulnerability mass. This captures a simple but operationally relevant fact: when the system is already near the frontier (x high), there is less new information to be produced for the ecosystem; when it is fragile (x low), informative refusal behavior can have outsized external value.

Intertemporal tradeoffs and discounting. Both the platform and the planner evaluate streams of payoffs with discount factor $\beta \in (0, 1)$. The platform chooses a policy for a_t to trade off immediate consequences (including any leakage cost $\Delta_E - \Delta_R$ and any private benefit $b_E - b_R$) against the dynamic effect of faster movement in x_t under higher α_E . The planner faces the same dynamic tradeoff but additionally internalizes the spillover value g .

Scope and limitations of the reduced form. This setup abstracts away many details: heterogeneous users (benign versus adversarial), multi-dimensional vulnerability surfaces, and explicit strategic adaptation by attackers. Those forces are partially folded into \bar{h} , Δ_a , and ω . The payoff is tractability: we obtain a one-dimensional state whose evolution is directly

controlled by refusal style, which lets us ask a sharp question that is difficult to answer qualitatively in deployment debates—namely, *when* should a platform be explanatory versus evasive, and how do institutional incentives shift that decision. In the next section we make this precise by defining the platform’s Markov decision problem and characterizing the resulting equilibrium policy.

4 Equilibrium concept: the platform’s Markov decision problem

Because refusal style is chosen by a single platform facing an evolving robustness state, the equilibrium object in our reduced form is a stationary Markov policy (often called a Markov-perfect policy by analogy to dynamic games, though there is no strategic opponent explicitly modeled here). A (stationary) Markov policy is a measurable map $\sigma : [0, 1] \rightarrow \{E, R\}$ that selects the refusal style as a function of the current robustness stock x . Given σ , the induced state process is $x_{t+1} = f_{\sigma(x_t)}(x_t)$ and the platform evaluates the discounted return

$$\mathbb{E} \left[\sum_{t \geq 0} \beta^t \pi_{\sigma(x_t)}(x_t) \mid x_0 = x \right], \quad (7)$$

where expectation is trivial in our deterministic transition but is useful notation for later extensions (e.g., stochastic incidents or noisy discovery).

Bellman equation and the dynamic programming operator. Let $V(x)$ denote the platform’s optimal value starting from state x . Standard dynamic programming yields the Bellman equation

$$V(x) = \max \left\{ \pi_E(x) + \beta V(f_E(x)), \pi_R(x) + \beta V(f_R(x)) \right\}. \quad (8)$$

Define the Bellman operator T acting on bounded functions $v : [0, 1] \rightarrow \mathbb{R}$ by

$$(Tv)(x) := \max_{a \in \{E, R\}} \left\{ \pi_a(x) + \beta v(f_a(x)) \right\}. \quad (9)$$

The equilibrium value function is a fixed point $V = TV$, and an optimal Markov policy can be recovered by selecting, for each x , an action attaining the maximum in (8). We will write the associated action-value functions as

$$Q_a(x; v) := \pi_a(x) + \beta v(f_a(x)), \quad (10)$$

so that $Tv(x) = \max\{Q_E(x; v), Q_R(x; v)\}$.

Existence and uniqueness of the value function. Our primitives imply a well-posed discounted Markov decision problem on a compact state space. First, payoffs are bounded on $[0, 1]$ because $\pi_a(x)$ is affine in x with finite coefficients. Second, the transition maps f_E, f_R are continuous and map $[0, 1]$ into itself.¹ Under these conditions, T is a contraction on the complete metric space of bounded functions with the sup norm: for any v, w ,

$$\|Tv - Tw\|_\infty \leq \beta \|v - w\|_\infty, \quad (11)$$

since $|\max_i u_i - \max_i \tilde{u}_i| \leq \max_i |u_i - \tilde{u}_i|$ and the only dependence on v is through $\beta v(f_a(x))$. By the Banach fixed-point theorem, there exists a unique bounded fixed point V solving (8), and value iteration $v^{(n+1)} = Tv^{(n)}$ converges uniformly to V from any bounded initial $v^{(0)}$. This uniqueness is about the *value* V ; optimal actions need not be unique at states where the platform is indifferent between E and R .

Monotonicity of the value function and comparative statics intuition. We will maintain (and later verify in examples) that V is increasing in x . Intuitively, higher x reduces baseline expected harm in every future period and, because the transition is increasing in x , it also leads to (weakly) higher future robustness under any fixed action sequence. Formally, if v is increasing, then Tv is increasing: each $Q_a(\cdot; v)$ is increasing because $\pi_a(x)$ is increasing in x (higher robustness reduces harm) and $f_a(x)$ is increasing with v increasing. Since value iteration preserves monotonicity and converges to V , it follows that V is increasing whenever we start from an increasing $v^{(0)}$ (e.g., $v^{(0)} \equiv 0$) and the monotonicity-preserving conditions hold. This property is central for the threshold characterization: it ensures that a larger learning step in x is always weakly more valuable in continuation terms.

From Bellman optimality to a threshold rule. To characterize the optimal policy, consider the action advantage

$$D(x) := \left(\pi_E(x) - \pi_R(x) \right) + \beta \left(V(f_E(x)) - V(f_R(x)) \right). \quad (12)$$

Choosing E is optimal at x if and only if $D(x) \geq 0$. Our goal is to establish a single-crossing property: $D(x)$ is (weakly) decreasing in x . If so, the set $\{x : D(x) \geq 0\}$ is an interval of the form $[0, x^*]$ (possibly degenerate), which implies a threshold policy.

The economic content of the single-crossing condition is straightforward. The current-period term $\pi_E(x) - \pi_R(x)$ captures the immediate net benefit of being explanatory rather than evasive (including any leakage penalty). The continuation term depends on how much more robustness we expect

¹Indeed, $f_a(x) = (1 - \omega)[(1 - \alpha_a)x + \alpha_a] \in [0, 1 - \omega] \subset [0, 1]$ for $\alpha_a \in [0, 1]$ and $\omega \in [0, 1)$.

tomorrow under E than under R , scaled by the marginal value of robustness encoded in V . Crucially, the incremental learning advantage

$$f_E(x) - f_R(x) = (1 - \omega)(\alpha_E - \alpha_R)(1 - x) \quad (13)$$

shrinks as x rises: when little vulnerability mass remains, there is less left for explanation to uncover, so the dynamic value of explanation attenuates.

To make this formal, we impose sufficient conditions that deliver decreasing differences. First, we assume $\pi_E(x) - \pi_R(x)$ is weakly decreasing in x .² Second, we use that $f_E(x) - f_R(x)$ is strictly decreasing in x , as shown above. Third, we require that V is increasing, and (for a convenient sufficient condition) that V is convex so that the mapping $y \mapsto V(y)$ exhibits increasing marginal value of y .³

Under these assumptions, $D(x)$ is weakly decreasing, hence the optimal policy is a threshold: there exists $x_D^* \in [0, 1]$ such that the platform chooses E if $x \leq x_D^*$ and R otherwise. When $D(x)$ is strictly decreasing (for example, when the continuation term is strictly decreasing and indifference occurs at most at a point), the threshold is essentially unique: any two optimal Markov policies can differ only on a set of states where $D(x) = 0$, which in the strict case is at most a singleton.

Indifference, tie-breaking, and (non-)uniqueness of policies. Even with a unique value function, the optimal action correspondence can be set-valued at indifference states. This matters operationally: two platforms with identical primitives could implement different refusal styles near the switching region if their evaluation or governance processes impose different tie-breaking rules (e.g., “prefer safer in the short run” vs. “prefer more informative outputs”). In our analysis, the main comparative statics and welfare conclusions depend on the location of the switching region rather than on behavior at a single knife-edge state, so we treat the threshold as unique up to these indifferences. Formally, if $D(x) > 0$ on an interval, E is uniquely optimal there; if $D(x) < 0$ on an interval, R is uniquely optimal there; if $D(x) = 0$ at isolated points, any selection is optimal.

Computability and verification perspective. The contraction property implies that the threshold policy can be computed by value iteration

²In the baseline payoff specification $\pi_a(x) = b_a - \kappa(\bar{h}(1 - x) + \Delta_a)$, the difference $\pi_E(x) - \pi_R(x) = (b_E - b_R) - \kappa(\Delta_E - \Delta_R)$ is constant in x , satisfying this assumption with equality. We state the condition in a slightly more general form to accommodate extensions where, e.g., explanation has a larger immediate leakage penalty in more fragile states.

³Convexity is not logically necessary for a threshold result but provides a clean route via monotone comparative statics: when V is convex and $f_E - f_R$ decreases in x , the difference $V(f_E(x)) - V(f_R(x))$ inherits a decreasing pattern. One can alternatively work with weaker curvature bounds or use a supermodularity argument in (x, a) under appropriately ordered transitions.

on a grid over $[0, 1]$, and verified by checking the sign of $D(x)$ (or the discrete analogue). This computational angle is more than a technicality: in deployment settings, refusal policies are often governed by measurable targets (incident rates, audit findings closed, evaluation scores), and a scalar index like x is a plausible abstraction of internal dashboards. The threshold structure says that, under diminishing informational returns to explanation, optimal behavior can be summarized by a simple rule: explain while the system is still meaningfully learnable, then become more evasive once marginal discovery falls below its immediate risk and product-cost tradeoff. In the next section we leverage this structure to compare private and social objectives when discovery has spillover value.

5 Main results: externalities, a closed-form switching rule, and welfare comparisons

Our central welfare claim is that refusal style is not merely a UX choice: it is an *information-production decision* that governs how quickly the robustness stock improves, and those informational gains are partly non-rival. In deployment, explanatory refusals can surface actionable details for internal triage, generate structured telemetry for auditors, and create reusable safety knowledge (e.g., red-team corpora, mitigations, and classifier features). These are precisely the kinds of benefits that a single platform may not fully internalize when its objective is limited to product benefit minus expected harm. We formalize this as a wedge between the platform’s and the planner’s switching thresholds.

Theorem 5.1 (Refusal Externality / private underprovision of explanation). *Suppose the single-crossing conditions in the enclosing scope hold so that both the platform and the planner admit (possibly degenerate) threshold-optimal Markov policies. If the planner’s per-period payoff adds spillovers*

$$g \cdot (\alpha_a - \alpha_R)(1 - x), \quad g \geq 0,$$

then the planner’s explanatory region weakly contains the platform’s:

$$x_D^* \leq x_W^*,$$

with strict inequality whenever $g > 0$ and the switching threshold is interior.

Proof sketch and interpretation. Let $D(x)$ denote the platform’s advantage of E over R at state x as in (12). The planner’s corresponding advantage is

$$\tilde{D}(x) = \left(\tilde{\pi}_E(x) - \tilde{\pi}_R(x) \right) + \beta \left(W(f_E(x)) - W(f_R(x)) \right) = D(x) + g(\alpha_E - \alpha_R)(1 - x) + \beta \left([W - V] \circ f_E - [W - V] \circ f_R \right)$$

The key observation is that, holding fixed the continuation values, the planner enjoys an additional *current* gain from choosing E at any $x < 1$:

$$\tilde{\pi}_E(x) - \tilde{\pi}_R(x) = \pi_E(x) - \pi_R(x) + g(\alpha_E - \alpha_R)(1 - x) \geq \pi_E(x) - \pi_R(x),$$

with strict inequality if $g > 0$ and $x < 1$. Under the maintained threshold structure (single-crossing), adding a nonnegative term that is decreasing in x shifts the indifference point weakly to the right. Operationally: even if explanatory refusals are privately unattractive once the system is already robust, the planner keeps explaining longer because the *marginal* informational return—though diminishing in x —still benefits third parties. The theorem thus identifies a concrete externality channel: the platform behaves *too evasively* relative to the social optimum.

A tractable switching rule under an “ R -continuation” approximation. While Theorem 5.1 is qualitative, governance often requires a quantitative rule-of-thumb: when is a temporary push toward explanation justified given a prevailing operational default of evasiveness? A simple and informative approximation is the one-step deviation test: compare (i) choose E today and then revert to R forever against (ii) choose R forever. This corresponds to evaluating whether a *single* explanatory intervention is worthwhile given a fixed operational continuation.

Let $V^R(x)$ denote the value under the stationary policy that always chooses R . Because payoffs are affine in x and the transition $f_R(x)$ is affine, V^R is affine:

$$V^R(x) = A_R + B_R x.$$

Write $f_R(x) = c_R + d_R x$ with $d_R = (1 - \omega)(1 - \alpha_R)$ and $c_R = (1 - \omega)\alpha_R$. Also write $\pi_R(x) = \bar{\pi}_R + \kappa \bar{h} x$, where $\bar{\pi}_R := b_R - \kappa(\bar{h} + \Delta_R)$. Substituting into the fixed-point equation $V^R(x) = \pi_R(x) + \beta V^R(f_R(x))$ yields the slope

$$B_R = \kappa \bar{h} + \beta B_R d_R \implies B_R = \frac{\kappa \bar{h}}{1 - \beta(1 - \omega)(1 - \alpha_R)} > 0. \quad (14)$$

The sign $B_R > 0$ formalizes the “shadow value” of robustness: improving x reduces baseline harm forever, discounted and attenuated by obsolescence and imperfect learning under R .

Now compare $\pi_E(x) + \beta V^R(f_E(x))$ to $V^R(x)$. Using linearity, the gain from a one-period switch to E is

$$\pi_E(x) - \pi_R(x) + \beta B_R(f_E(x) - f_R(x)) = (b_E - b_R) - \kappa(\Delta_E - \Delta_R) + \beta B_R(1 - \omega)(\alpha_E - \alpha_R)(1 - x). \quad (15)$$

Thus, under the R -continuation test, choosing E is optimal if and only if (15) is nonnegative. The implied threshold is

$$x \leq x^\dagger := 1 - \frac{\kappa(\Delta_E - \Delta_R) - (b_E - b_R)}{\beta B_R(1 - \omega)(\alpha_E - \alpha_R)}, \quad (16)$$

clipped to $[0, 1]$ when the numerator is negative (then $x^\dagger \geq 1$, so E is always worthwhile under the test) or when the numerator is large (then $x^\dagger \leq 0$, so E is never worthwhile under the test).

Two aspects of (16) matter for mechanism design. First, the dynamic term scales with $(1 - x)$: explanation is most valuable when the system is still far from fully robust, consistent with the diminishing-returns intuition behind the global threshold structure. Second, the multiplier βB_R converts “faster learning” into “future harm reduction”; by (14), this shadow value is larger when baseline harm severity \bar{h} is high, when the decision-maker is more patient (high β), and when robustness is persistent (low ω and/or low α_R so that R does not already close vulnerabilities quickly).

Welfare comparison and a governance-relevant wedge. Theorem 5.1 implies that the private policy induces a lower steady-state rate of vulnerability discovery than is socially optimal whenever $g > 0$ and x is not already near 1. In welfare terms, let σ_D and σ_W denote the platform’s and planner’s optimal threshold policies, and let x_t^D, x_t^W be the induced state trajectories from a common x_0 . The welfare gap can be written as

$$W(x_0) - V(x_0) = \sum_{t \geq 0} \beta^t \left(\tilde{\pi}_{\sigma_W(x_t^W)}(x_t^W) - \pi_{\sigma_D(x_t^D)}(x_t^D) \right),$$

which combines (i) direct spillovers $g(\alpha_E - \alpha_R)(1 - x)$ in periods where the planner explains and the platform does not, and (ii) the downstream effect that earlier explanation raises future x and thereby reduces baseline harm $\kappa \bar{h}(1 - x)$ in all subsequent periods. Even if one is agnostic about the precise magnitude of g , the model isolates a practical failure mode: platforms can rationally select evasive refusals because they internalize immediate leakage risk $\Delta_E - \Delta_R$ but not the ecosystem-level returns to producing structured information about what the model would have done.

Finally, it is important to flag a boundary case: if $\kappa(\Delta_E - \Delta_R)$ is large enough relative to both $(b_E - b_R)$ and the dynamic value in (15), then *both* the platform and the planner optimally choose evasiveness everywhere. This is the regime where explanation cannot be made safe at the margin (e.g., because it reliably leaks operationally useful details to attackers). In such cases, the externality result does not argue for unconditional mandates; rather, it highlights an R&D target: reduce Δ_E (through safer templating, constrained decoding, or redaction) and/or increase $\alpha_E - \alpha_R$ (through better monitoring pipelines) to move the system into a region where explanatory refusal is dynamically and socially beneficial. The next section turns to comparative statics that make these levers explicit and clarify when standards should push platforms toward explanation.

6 Comparative statics and policy levers

The threshold characterizations above let us read comparative statics directly as statements about when explanatory refusal is worth its (possibly) higher immediate leakage risk. Under the tractable R -continuation rule (16), the explanatory region expands whenever the dynamic term

$$\beta B_R(1 - \omega)(\alpha_E - \alpha_R)(1 - x)$$

becomes large relative to the immediate net cost $\kappa(\Delta_E - \Delta_R) - (b_E - b_R)$. This decomposition is useful because each factor corresponds to a concrete deployment or governance lever: patience and horizon (β), the shadow value of robustness (B_R , driven by harm severity and persistence), the speed at which robustness decays (ω), and the incremental informativeness of explanation ($\alpha_E - \alpha_R$). In what follows, we interpret these levers as audit design choices, threat-model properties, and evaluation/regulatory incentives, and we spell out when standards should *mandate* explanation versus when they should merely *enable* it.

Audit intensity and instrumentation: raising α_R can reduce the marginal value of explanation. A common argument for explanatory refusals is that they create better feedback for internal safety work. But in practice, a large fraction of feedback comes from telemetry (prompt logs, classifier scores, sandboxed tool traces, and post-hoc incident analysis) rather than from the literal refusal text. In our reduced form, better auditing and instrumentation can be modeled as increasing the discovery rate even under evasiveness, i.e. raising α_R . This has two opposing implications.

First, increasing α_R shrinks the gap $\alpha_E - \alpha_R$, reducing the incremental learning advantage of explanation and thus lowering the threshold at which E is privately optimal. This is the “substitution” effect: if we can learn almost as well from evasive refusals (because audits are strong), then explanatory text adds less unique value.

Second, stronger audits can *also* change the immediate-risk terms Δ_E, Δ_R . For example, templated explanations with automated redaction, plus continuous monitoring for prompt-injection and policy circumvention, can reduce the effective leakage risk of explanation, lowering Δ_E without commensurately lowering α_E . This is a “complementarity” effect: audit systems make explanation safer *and* more actionable (by tying refusals to structured labels, reproduction harnesses, and mitigation pipelines). Which effect dominates is empirical and depends on whether the organization is telemetry-limited (low α_R) or mitigation-limited (high Δ_E due to uncontrolled phrasing and unconstrained decoding).

Governance implication: standards that only demand “more auditing” can unintentionally make mandated explanation less attractive by pushing

α_R up without reducing Δ_E . If the goal is to elicit socially valuable explanation, audit requirements should be paired with *safe explanation protocols* (e.g. fixed refusal templates, constrained generation, and redaction rules) that specifically target Δ_E .

Attacker innovation / obsolescence ω : continual change favors longer explanatory phases, up to a point. The obsolescence parameter ω captures the intuition that robustness is not a one-time capital stock: new attack methods appear, distribution shift occurs, and mitigations decay. In the transition $f_a(x) = (1 - \omega)[(1 - \alpha_a)x + \alpha_a]$, higher ω directly attenuates the carry-over of today’s robustness into tomorrow. This creates two forces.

On the one hand, higher ω raises the *need* for ongoing discovery: without continual learning, x erodes mechanically. In a richer model where E could be chosen repeatedly, this tends to push optimal policies toward more frequent information production, because the system must “run to stay in place.”

On the other hand, in the one-step expression (15), ω scales down the immediate effect of learning on next period’s robustness through the factor $(1 - \omega)$. Thus, extreme obsolescence reduces the effectiveness of *any* one-period discovery intervention. Put differently: when the world changes too quickly, explanation may still be socially valuable for ecosystem learning, but the platform’s private incentive to invest in robustness through explanation can weaken because the platform cannot reliably bank the gains.

Governance implication: when ω is high (rapidly evolving threat landscape), standards should emphasize *continuous* monitoring and mitigation pipelines rather than sporadic disclosure; if explanation is mandated, it should be coupled to rapid patch loops so that the information produced can be converted into durable improvements before obsolescence erases them.

Harm severity and liability: $\kappa\bar{h}$ increases the shadow value B_R and can make explanation privately rational. Equation (14) shows B_R scales linearly with $\kappa\bar{h}$. Interpreting κ as a per-unit social/penalty cost and \bar{h} as baseline harm intensity, their product captures the extent to which residual vulnerability mass $(1 - x)$ is costly. When harms are severe (e.g. biosecurity, scalable cyber abuse) or liability is salient (expected penalties, contractual damages, or enforcement risk), the shadow value of moving x upward increases. Since B_R enters the dynamic benefit term in (15), higher $\kappa\bar{h}$ expands the parameter region where E is optimal even if explanation slightly increases immediate risk (moderate $\Delta_E - \Delta_R$).

This is an important corrective to a common intuition: “if harms are severe, we should be more evasive.” Our model says the opposite can hold dynamically: severe harms raise the value of accelerating robustness, which can justify explanatory refusal *provided* the explanation can be made safe enough (i.e. Δ_E is controllable). The relevant tradeoff is not “explain vs be

safe,” but “invest in discoverability vs accept persistent risk.”

Governance implication: in high-severity domains, standards should not treat evasiveness as the default safe harbor. Instead, they should specify conditions under which explanation is required (for learnability) and conditions under which it is prohibited (when Δ_E cannot be bounded), with an explicit burden on developers to demonstrate control of Δ_E via redaction and constrained decoding.

Evaluation penalties for evasiveness: shifting $b_E - b_R$ can close the private–social wedge without requiring implausible altruism. The term $(b_E - b_R)$ captures private benefits of non-evasiveness: user trust, product usefulness, and procurement/evaluation rewards for being forthright. Evaluators and regulators can operationalize this by penalizing evasive refusals (lowering b_R) or rewarding safe, structured explanations (raising b_E). Because $(b_E - b_R)$ enters the switching condition linearly, even modest evaluation incentives can substantially shift the threshold when the learning term is already positive.

However, there is a failure mode: rewarding “explanation” without specifying safety constraints can increase Δ_E by inducing models to be more verbose, more specific, or more negotiable in refusal contexts. This is a mechanistic Goodhart problem: optimizing for explanation quality can push systems toward revealing actionable detail unless the explanation format is constrained.

Governance implication: evaluation should score *structured* explanations that are (i) non-actionable, (ii) consistently templated, and (iii) instrumented (linkable to internal taxonomy and mitigation tickets). In our terms, the goal is to raise $b_E - b_R$ while simultaneously lowering $\Delta_E - \Delta_R$.

When should standards mandate explanation? A thresholded mandate tied to measurable proxies. The externality result implies that when $g > 0$ there are states where the planner prefers E but the platform prefers R . A practical standard cannot observe x directly, but it can mandate explanation contingent on proxies for “how much is left to discover.” Examples include: recent incident rate, red-team yield (fraction of novel findings), distribution-shift indicators, and measured robustness scores on adversarial suites.

A conservative rule consistent with the model is: mandate explanatory refusal in regimes where (a) residual risk is high (low proxy for x), (b) the incremental learning advantage is real (large measured $\alpha_E - \alpha_R$ under the organization’s telemetry), and (c) the incremental leakage risk is demonstrably bounded (small $\Delta_E - \Delta_R$ under controlled evaluations). Conversely, when Δ_E cannot be controlled (e.g. explanations reliably leak attacker-relevant details), standards should *not* mandate explanation; they should mandate

investments that move the system into the explainable-safe regime, such as constrained decoding, redaction, and better post-deployment monitoring.

Framed this way, the mandate is not ideological (“always explain”) but state-dependent: *explain while fragile, evade when robust or when explanation is unsafe*. The model’s comparative statics then become a checklist for policy design: raise $b_E - b_R$ via evaluation incentives, reduce Δ_E via safe templating, increase $\alpha_E - \alpha_R$ via better labeling and triage pipelines, and treat high- ω domains as requiring continuous rather than episodic information production.

7 Extensions (optional, tractable)

The baseline model deliberately collapses several real deployment details into a single robustness state x and a binary refusal style. That abstraction is useful for isolating the dynamic tradeoff, but we can extend it in ways that (i) remain analytically close to the threshold logic above, and (ii) map more directly to operational decisions (rate-limiting, user segmentation, safety taxonomy design, and controlled degrees of disclosure). We sketch four such extensions; in each case, the key question is whether we preserve a monotone “explain-when-fragile” structure, and what new failure modes appear.

(i) Endogenous attack frequency and strategic traffic shaping. A missing channel in the reduced-form payoff is that refusal behavior can change how often risky prompts are attempted. If explanations are perceived as informative or negotiable, they may increase probing and adaptation; conversely, they may deter low-effort attempts by making policy boundaries clear. A tractable way to capture this is to let the arrival rate of risky prompts be $m(a, x) \geq 0$, so that per-period harm and benefit scale with traffic:

$$\pi_a(x) = m(a, x) \left(b_a - \kappa(\bar{h}(1 - x) + \Delta_a) \right), \quad x_{t+1} = f_a(x_t).$$

If $m(E, x) > m(R, x)$ for low x (explanations invite more probing when the system is fragile), then explanation becomes less attractive precisely where learning benefits are highest, creating a sharper tradeoff. However, the extension remains tractable if $m(a, x)$ is decreasing in x (robust systems attract fewer successful probes) and if the difference $m(E, x) - m(R, x)$ is itself decreasing in x (probing externalities are largest when fragility is obvious). Under these monotonicity conditions, the action-value difference can still exhibit single-crossing, but the threshold generally shifts downward.

This formulation also makes explicit a governance-relevant lever: traffic shaping. Rate limits, friction, and abuse monitoring can be modeled as reducing $m(a, x)$ without directly changing f_a . In practice, this means we

can sometimes “buy back” explanatory refusals by coupling them to anti-probing controls, rather than treating explanation as intrinsically unsafe.

(ii) Heterogeneous users: benign demand vs malicious probing.

Our baseline payoff b_a implicitly mixes benign user value with any reputational or product effects of refusal style. A more explicit decomposition separates benign and malicious prompt streams with different objective weights. Let $\lambda \in [0, 1]$ be the share of risky-prompt mass that is malicious, with $(1 - \lambda)$ benign-but-risky (e.g., users requesting disallowed content without adversarial intent). Suppose the developer derives benefit b_a^B from benign interactions and faces harm from malicious interactions. A simple specification is

$$\pi_a(x) = (1 - \lambda)b_a^B - \lambda \kappa(\bar{h}(1 - x) + \Delta_a^M) - (1 - \lambda)\kappa\Delta_a^B,$$

where Δ_a^M captures incremental immediate misuse risk and Δ_a^B captures benign-side costs (e.g., chilling effects, user frustration) that we may want to penalize as welfare loss. The state transition can still be driven by aggregate discovery, but we can now allow discovery to depend on user type: explanations may elicit more actionable reports from benign users (raising effective learning) while also improving malicious adaptation (raising immediate risk). One tractable reduced form is to keep $x_{t+1} = f_a(x_t)$ with α_a interpreted as net discovery after internal triage, while allowing Δ_a^M to scale with λ .

Two qualitative implications follow. First, policies may become *segmented*: we can justify explanatory refusals for authenticated, low-risk cohorts (small effective λ) while remaining evasive for untrusted traffic. Second, the “threshold in x ” may interact with a “threshold in λ ”: explanation is optimal when the system is fragile *and* the audience is sufficiently benign. This resonates with common deployment practice (graduated access, KYC, research programs) and clarifies an alignment failure mode: if user-mix shifts over time (e.g., during a viral event), a policy calibrated for low λ can become unsafe without any change in x .

(iii) Multi-category harms with different learning rates and obsolescence. Real systems face multiple hazard classes (e.g., cyber, fraud, self-harm, bio) with different mitigation pipelines, different rates of attacker adaptation, and different externalities. A tractable extension replaces scalar x with a vector $x = (x^{(1)}, \dots, x^{(K)})$, where each component tracks mitigated mass for category k . Let the transition be separable:

$$x_{t+1}^{(k)} = f_a^{(k)}(x_t^{(k)}) = (1 - \omega_k)[(1 - \alpha_{a,k})x_t^{(k)} + \alpha_{a,k}],$$

and let per-period harm aggregate additively with category weights $\kappa_k \bar{h}_k$:

$$\pi_a(x) = b_a - \sum_{k=1}^K \kappa_k (\bar{h}_k(1 - x^{(k)}) + \Delta_{a,k}).$$

This remains computationally simple (affine dynamics and linear payoffs), but introduces a choice: the action a is shared across categories, while learning and leakage effects vary by category. If category-level terms differ sharply (e.g., $\Delta_{E,k}$ is large for bio but small for cyber, while $\alpha_{E,k} - \alpha_{R,k}$ is the opposite), then a single global refusal style is inherently mis-specified. One response is to allow category-contingent actions $a^{(k)} \in \{E, R\}$, which yields K independent scalar problems and hence K thresholds. Another response is to retain a global action but interpret it as a *policy bundle* whose effective parameters are weighted averages induced by classifier routing and templating. Either way, the extension makes a concrete empirical prediction: category suites with higher ω_k (faster attacker innovation) should exhibit longer “explanatory phases” provided $\Delta_{E,k}$ can be bounded, whereas categories with high irreversible leakage risk should remain evasive even when fragile.

A further open problem is combinatorial: if we add a constraint on the *total* explanation budget (e.g., only a fraction of refusals can be explanatory due to reviewer capacity), we obtain a knapsack-like dynamic allocation problem. The affine structure suggests index-style heuristics (e.g., explain in the category with highest marginal shadow value of $x^{(k)}$), but formal optimality generally requires stronger assumptions.

(iv) Partial explanations: continuous disclosure and safe templates.

Finally, the binary action set is too coarse for many safety protocols. In practice we choose how much to disclose: a terse refusal, a generic policy citation, a high-level harm rationale, or a detailed (but redacted) explanation with pointers to safe alternatives. We can model this with a continuous action $e \in [0, 1]$ representing explanation intensity. A tractable parameterization is

$$\alpha(e) = \alpha_R + e(\alpha_E - \alpha_R), \quad \Delta(e) = \Delta_R + e(\Delta_E - \Delta_R), \quad b(e) = b_R + e(b_E - b_R),$$

yielding the same affine transition $x_{t+1} = f_{\alpha(e)}(x_t)$ and linear payoffs in e . With this linear mixing, the per-period objective is linear in e plus a continuation term that is concave/convex depending on V . Under mild convexity of V , the optimum often remains “bang-bang” (choose $e = 0$ or $e = 1$), which rationalizes why organizations gravitate toward discrete refusal templates. But if we add diminishing returns to explanation (concave $\alpha(e)$) or sharply convex leakage risk (convex $\Delta(e)$), interior solutions emerge: the optimal policy uses *partial* explanation when x is moderate, reserving full explanation for the most fragile regimes.

This extension also cleanly represents “safe explanation protocols” as constraints on the feasible set: we can cap $\Delta(e)$ by enforcing templated language, constrained decoding, and redaction, effectively allowing higher $\alpha(e)$ at lower $\Delta(e)$. In other words, protocol design shifts the action frontier rather than merely picking a point on it.

Across these extensions, the overarching methodological point is that we can preserve the core threshold intuition while adding operational degrees

of freedom. Doing so sets up the next step: specifying how to measure the effective α 's, Δ 's, and ω 's from deployment data and evaluations, so that the model's state-dependent prescriptions become testable rather than rhetorical.

8 Empirical design sketch: measuring discovery and adaptive robustness

To make the dynamic tradeoff actionable, we need to operationalize three latent objects that our theory treats as primitives: (i) the robustness state x_t (how much vulnerability mass is already “found and mitigated”), (ii) the action-dependent discovery rates α_E, α_R (how quickly remaining mass is converted into mitigations under different refusal styles), and (iii) the obsolescence rate ω (how quickly newly relevant vulnerability mass appears, or how quickly old mitigations stop working). The empirical challenge is that none of these are directly observed: what we see are noisy incidents, evaluation scores, red-team findings, and policy outputs, all under strategic adaptation and shifting traffic.

Measuring the state x_t via a latent robustness index. A practical proxy for x_t is a severity-weighted “residual risk” score built from standardized adversarial evaluations. Suppose we maintain a fixed evaluation battery of unsafe tasks (plus periodically refreshed tasks to detect overfitting) and record a success probability s_t (higher means more successful jailbreaks/misuse). A simple mapping is $x_t \approx 1 - s_t$, but this conflates category mix and changes in evaluator difficulty. A more robust approach is to treat x_t as a latent factor estimated from multiple signals: (a) jailbreak success on several suites (public, internal, third-party), (b) severity-weighted open vulnerability backlog (e.g., number of unresolved high-severity policy bypasses), (c) incident rate among monitored abuse channels, and (d) time-to-mitigation for newly discovered issues. Formally, we can use a state-space model with observation equations of the form

$$y_t^{(j)} = c_j + d_j x_t + \nu_t^{(j)},$$

where $y_t^{(j)}$ are logged metrics (transformed to approximate linearity) and x_t follows the controlled transition $x_{t+1} = f_{a_t}(x_t) + \epsilon_t$. This allows us to filter x_t over time (Kalman-style if Gaussian, particle filtering otherwise) and to quantify uncertainty, which matters because policy thresholds should be risk-adjusted rather than point-estimated.

Operationalizing “discovery rate” α_a from mitigation kinetics. In the model, α_a is not merely the rate at which we *observe* new failure modes;

it is the effective conversion of remaining vulnerability mass into durable mitigations after triage. Empirically, we can approximate α_a by tracking the flow of *actionable* discoveries that lead to model or system changes (training data additions, rule updates, tool hardening) and then verifying that those changes reduce measured residual risk. Concretely, we log a pipeline: (i) candidate failures (user reports, automated monitors, red-team probes), (ii) validated failures, (iii) mitigations deployed, and (iv) post-mitigation regression results on holdout attack sets. The key is to discount discoveries that are duplicates or non-generalizing. If D_t is the severity-weighted count of *novel* validated vulnerabilities at time t , and \hat{x}_t is our filtered robustness estimate, then a reduced-form estimator consistent with the transition structure is obtained by fitting

$$\hat{x}_{t+1} = (1 - \omega)[(1 - \alpha_{a_t})\hat{x}_t + \alpha_{a_t}] + \eta_t$$

by maximum likelihood (with α_{a_t} taking one of two values depending on whether the deployed refusal policy is predominantly E or R). In practice, we will want to allow for partial compliance (mixed templates, model variance), so we can replace the binary a_t with the observed fraction of explanatory refusals in period t and estimate an “effective” α as a function of that fraction.

Estimating obsolescence ω under nonstationarity. The term ω captures adversary innovation and concept drift: even if we stop changing the model, the environment changes. We can estimate ω by observing degradation in robustness metrics during intervals with minimal mitigation updates (or by controlling for update size). For example, if we have a “frozen” evaluation harness (or a stable subset) and we observe $\hat{x}_{t+1} < \hat{x}_t$ despite no major interventions, that decay is informative about ω .⁴ More realistically, we fit ω jointly with α_a in the state-space model, letting ω vary by category or by threat surface when the multi-category extension is used.

Training regimes as interventions on $(b_a, \Delta_a, \alpha_a)$. A core implication of the theory is that refusal style is a policy choice whose value depends on how it shifts immediate harm Δ_a and discovery α_a , as well as user-facing utility b_a . Different training regimes move these parameters differently. For example, reward-model training that explicitly values “helpful refusals” can increase $b_E - b_R$ (users prefer non-evasive responses) and may also increase $\alpha_E - \alpha_R$ if explanations yield higher-quality feedback and debugging signals. However, the same training can increase $\Delta_E - \Delta_R$ if explanations leak procedural details that enable adaptation. This suggests a disciplined experimentation loop: define a family of refusal templates or decoding constraints

⁴This is imperfect because absence of logged interventions does not imply absence of effective changes (e.g., upstream model updates, tooling modifications). The identification strategy should therefore instrument for true update intensity.

(the treatment), measure how they change leakage proxies and discovery yield, and then compute the implied threshold shift. Importantly, we should treat “explanatory refusal” as a *bundle* (templating, redaction, constrained decoding, reference to safe alternatives), because Δ_E is largely a function of protocol design rather than a fixed property of being “explanatory.”

Outcome metrics: immediate harm, dynamic robustness, and spillovers.

We need metrics aligned to each term in the objective. Immediate harm proxies include: (i) success rate on high-risk tasks under a red-team harness, (ii) severity-weighted policy violation rate in monitored traffic, and (iii) downstream incident reports (fraud, malware generation, etc.) after appropriate attribution controls. Dynamic outcomes include: (a) time-to-detection of novel failures, (b) time-to-mitigation, (c) post-mitigation generalization (reduction in success on *held-out* attacks, not just the discovered prompt), and (d) stability under refreshed attacks (measuring effective ω). To capture the spillover value g , we can measure externalities such as (1) the rate of high-quality third-party vulnerability reports attributable to refusal explanations, (2) cross-org patch diffusion (e.g., whether shared taxonomies or disclosed failure modes reduce incidents elsewhere), and (3) ecosystem-level indicators like reduced prevalence of known attack patterns. None of these is perfect, but the point is to treat spillovers as measurable outputs rather than rhetorical claims.

Identification strategy: randomized rollout and quasi-experiments.

The central causal question is how switching from R to E (or increasing explanatory intensity) changes both near-term harm and subsequent robustness. The cleanest design is a randomized controlled rollout: randomize refusal style at the level of (i) user cohorts (authenticated research users versus baseline), (ii) geographic regions, or (iii) time blocks, with guardrails that cap worst-case harm. Because interference is plausible (attackers can adapt and share prompts), cluster randomization and careful monitoring are required. When randomization is infeasible, we can use quasi-experimental designs: difference-in-differences around policy changes, regression discontinuity when a threshold rule is introduced (e.g., explanation allowed only above a trust score), or instrumental variables using exogenous shifts in explanation feasibility (e.g., temporary staffing increases that improve triage speed and hence effective α without directly changing leakage). Across designs, we must explicitly model selection: if E changes the arrival rate of risky prompts, then naive comparisons confound policy effects with traffic composition. Logging prompt volumes, attacker novelty, and user trust signals is therefore part of the identification plan, not an afterthought.

A practical estimation loop. Putting these pieces together, we can implement a repeated cycle: (1) deploy a controlled variation in refusal policy (template, decoding constraints, or training regime); (2) estimate its effect on immediate harm metrics (informing $\Delta_E - \Delta_R$) and on discovery pipeline outputs (informing $\alpha_E - \alpha_R$); (3) update the filtered robustness state \hat{x}_t and re-estimate $(\alpha_E, \alpha_R, \omega)$ with uncertainty intervals; and (4) compute the implied privately optimal and socially optimal switching rules under candidate governance weights (κ, g) . The end goal is not to pretend the world is exactly affine-Markov, but to use the structure to force a consistent accounting of (i) what we gain dynamically from transparency and explanation, (ii) what we risk immediately, and (iii) where governance mechanisms must intervene when spillovers are large.

These empirical considerations set up the next step: translating the wedge between private and social thresholds into concrete policy levers (audits, procurement requirements, and transparency norms), while being honest about where the model’s tractability ends and numerical methods become unavoidable.

9 Discussion and policy implications: procurement, audits, and transparency

Our threshold results are not merely a stylized “behavioral” claim about tone; they imply that refusal style is an *intertemporal safety investment* with a predictable wedge between private and social incentives when discovery benefits spill over. That wedge is precisely what governance and organizational design are meant to address. In this section we translate the model into (i) procurement language that creates enforceable incentives, (ii) audit designs that measure the primitives our switching rule depends on, and (iii) transparency requirements that increase g -type spillover benefits while managing Δ_E -type leakage risk. We then flag where the tractable affine Markov structure ends and where numerical methods (and richer data) become unavoidable.

Procurement language: contracting on refusal *quality* rather than refusal *rate*. A recurring failure mode in practice is that buyers contract on a coarse scalar—e.g., “refuse all unsafe requests” or “maintain a policy violation rate below τ ”—which encourages evasive refusals that satisfy surface-level compliance while starving the monitoring pipeline of actionable signal. Our framework suggests contracting on a *profile* of refusal behavior: when the system is in a low-robustness regime (low x), buyers should want a higher share of “helpful-safe” refusals that generate diagnostic information and steer users to legitimate alternatives, provided leakage is controlled.

Concretely, procurement documents can require (a) a refusal taxonomy (e.g., categories of unsafe intent), (b) a calibrated “explanatory refusal” tem-

plate family with redaction rules, and (c) evidence that these templates increase effective discovery (a measurable proxy for $\alpha_E - \alpha_R$) without materially increasing immediate harm (a proxy for $\Delta_E - \Delta_R$). Language that is both enforceable and model-aligned includes clauses such as: (i) minimum coverage of explanatory refusals on *risky* prompts under an agreed classifier (with dispute resolution for misclassification), (ii) minimum “actionability” standards (refusals must cite policy category and provide safe alternatives, not generic deflection), and (iii) reporting obligations on mitigation kinetics (time-to-triage, time-to-fix, and post-fix regression performance). Importantly, buyers can allow the platform to be evasive in narrowly defined high-leakage categories (where Δ_E is demonstrably large), which avoids turning the model into a blanket pro-transparency slogan.

Audit design: estimating α -gaps and auditing the switching frontier. Audits that only test one-period harmfulness miss the dynamic quantity we care about: whether the system is *learning* from pressure. The operational audit target implied by the theory is the platform’s effective switching behavior: do we observe explanatory refusals concentrated when residual vulnerability is high (low \hat{x}_t), and do those refusals actually accelerate mitigation (higher estimated α) rather than merely sounding better?

A practical audit protocol can be organized around three linked checks. First, a *measurement audit* verifies the latent robustness index \hat{x}_t (or a proxy) is computed from a stable, versioned evaluation harness with documented refresh procedures and severity weights; auditors should be able to reproduce the index from logs and evaluation artifacts. Second, a *causal policy audit* estimates the effect of explanatory intensity on (i) short-run misuse success and (ii) downstream discovery yield, using randomized rollouts where feasible or quasi-experimental designs otherwise; the output is an audited estimate of $(\Delta_E - \Delta_R, \alpha_E - \alpha_R)$ with uncertainty intervals. Third, an *incentive-compatibility audit* checks whether the deployed policy is consistent with an announced decision rule (explicit threshold, or an equivalent risk score rule) and whether deviations are justified by documented changes in primitives (e.g., a newly discovered leakage channel that increases Δ_E). This is the governance analog of verifying that the system is following a stable Markov policy rather than opportunistically shifting to evasiveness when scrutiny rises.

Transparency requirements: increasing spillovers while bounding leakage. Because the planner’s advantage comes from the spillover term $g(\alpha_E - \alpha_R)(1 - x)$, governance can act either by increasing g (making learning benefits more socialized) or by requiring the private actor to internalize it (through liability, penalties, or mandated practices). Transparency is the most direct way to increase g , but it must be targeted: indiscriminate pub-

lication can also raise Δ_E by enabling faster attacker adaptation.

We therefore favor “structured transparency” requirements that separate *diagnostic* information from *procedural* information. Examples include: (i) publishing refusal taxonomies and high-level rationales (what category triggered the refusal) without disclosing operational details that would help jail-breaks; (ii) sharing de-identified aggregates on novel failure discovery and mitigation timing; (iii) maintaining a secure disclosure channel for vetted researchers, with time-bounded embargoes and standardized severity scoring; and (iv) releasing red-team datasets in delayed or filtered form to reduce immediate exploitability. The common theme is to turn refusal interactions into auditable safety signals without turning them into attacker training data. This is also where tooling matters: templating, constrained decoding, and automatic redaction can reduce Δ_E while preserving the explanatory content that drives α_E .

Regulatory posture: performance standards plus process standards. If a regulator attempts to mandate a single refusal style, it will fail in the corner cases our model already highlights: when explanatory content is unavoidably high-leakage, a strict explanation mandate can backfire. Instead, the model motivates a hybrid regime. Performance standards cap realized harm (bounding $\bar{h}(1 - x) + \Delta_a$ proxies), while process standards require demonstrable learning dynamics: versioned evaluations, incident reporting, and evidence that risky-prompt handling contributes to mitigation rather than simply suppressing outputs. In other words, we want to regulate both the level of risk and the *slope* of risk reduction over time.

Limits of the tractable model. The affine one-dimensional state is doing a lot of work. In deployment, vulnerability mass is multi-surface (bio, cyber, fraud, persuasion), and actions are richer than $\{E, R\}$: platforms can gate by user trust, throttle, rate-limit, watermark, route to tools, or request verification. Moreover, the environment is strategic: attacker effort responds to policy, which means ω and even the effective arrival rate of risky prompts are endogenous. Finally, partial observability is the norm: \hat{x}_t is filtered with noise, and the platform faces ambiguity about whether apparent improvements are robust or merely benchmark overfitting.

These limitations matter for interpretation. Our threshold result should be read as a *qualitative organizing principle* (single-crossing between immediate leakage risk and dynamic learning benefit), not as a claim that one scalar index is sufficient for all safety decisions. The empirical program sketched earlier is therefore not optional; it is what anchors the abstractions.

What requires numerical methods (and what to compute). The moment we move beyond the affine single-state setting, closed forms largely

disappear and we need numerical dynamic programming or simulation-based methods. Three extensions are especially important. First, a *multi-dimensional state* $x \in [0, 1]^K$ (by hazard category) with action-dependent cross-effects requires solving a higher-dimensional Bellman equation; approximate dynamic programming (e.g., fitted value iteration) or policy-gradient methods with safety constraints become relevant. Second, *stochastic discovery* and *rare catastrophic events* break the linear-quadratic intuition; we then want risk-sensitive objectives (CVaR, worst-case robust control) and constraint-based formulations, which typically require numerical solvers. Third, *partial observability* turns the problem into a POMDP: the sufficient statistic is a belief over x , updated via the audited observation model. Here, particle filters plus belief-state planning (or certainty-equivalent approximations with conservative margins) are the practical route.

From a governance perspective, the key point is that numerical methods do not weaken the policy message; they sharpen it. Once the primitives are estimated with uncertainty, the switching rule becomes a *distributional* statement (e.g., explain when $\Pr(x \leq x^*)$ is high enough), which is exactly what audits and procurement should anticipate: decisions under uncertainty, with documented risk margins and reproducible computation.

Taken together, procurement, audits, and transparency can be designed to make the socially valuable part of explanatory refusal legible and contractible, while bounding the leakage channels that would otherwise dominate. This sets the stage for our conclusion: the central challenge is not choosing between “helpful” and “safe,” but engineering refusal protocols and governance mechanisms so that helpfulness contributes to cumulative robustness rather than merely redistributing risk over time.

10 Conclusion

We set out to formalize a tension that practitioners routinely encounter but rarely articulate in intertemporal terms: the same refusal that reduces immediate misuse can either (i) accelerate the system’s discovery-and-mitigation pipeline by producing structured signal, or (ii) starve that pipeline by withholding any actionable information. In our model, this is the distinction between explanatory refusal E and evasive refusal R . The central contribution is to treat refusal style not as a static “tone choice,” but as a control variable in a discounted dynamic optimization problem, where today’s refusal affects tomorrow’s robustness through learning rates (α_E, α_R) and where the environment continuously regenerates residual risk through obsolescence ω .

Two qualitative results organize the analysis. First, under mild single-crossing conditions—the incremental learning advantage of explanation shrinks as robustness improves, and the immediate advantage of explanation does not grow with x —the platform’s optimal policy has threshold structure: ex-

plain when residual vulnerability is high (low x), and become more evasive as the system approaches maturity (high x). This matters because it turns an otherwise fuzzy governance debate (“should the model be more explanatory?”) into an auditable behavioral claim: the relevant object is the platform’s *switching frontier* in state space. In particular, the model implies that persistent, unconditional evasiveness is dynamically hard to justify whenever explanatory refusals measurably increase discovery and the decision-maker places nontrivial weight on future safety.

Second, when we add a spillover term capturing ecosystem learning—the idea that diagnostic refusal content can improve safety beyond the deploying platform (via shared benchmarks, external red-teaming, improved defender playbooks, or simply reduced systemic risk)—we obtain a refusal externality. The social planner values $g(\alpha_E - \alpha_R)(1 - x)$ in addition to the platform’s private payoff. This shifts the planner’s switching frontier outward: the socially optimal region for explanation is (weakly) larger, $x_D^* \leq x_W^*$, and strictly so when the spillover is positive and the threshold is interior. The practical interpretation is not that platforms “should always explain,” but that private incentives predictably underprovide explanation in precisely the regimes where marginal discovery is highest. This is the same structural logic that motivates safety case requirements, incident reporting, and shared vulnerability disclosure in other security domains: information production is socially valuable and privately undersupplied.

The policy relevance of the formalism is therefore less about advocating a particular refusal template and more about identifying what must be *made legible* to align incentives. The switching rule depends on a small set of primitives: the immediate net effect of explanation on harm and utility, $(b_E - b_R) - \kappa(\Delta_E - \Delta_R)$; the dynamic effect on robustness, scaled by $(\alpha_E - \alpha_R)$; the pace of obsolescence ω ; and the effective discount factor β (including any institutional analogs, such as long-term contracts, reputational capital, or liability regimes). Governance interventions map cleanly onto these objects: procurement can increase the private return to non-evasiveness (raising $b_E - b_R$) while bounding leakage (lowering $\Delta_E - \Delta_R$); auditing can estimate α -gaps and verify that the deployed policy responds to measured risk; and transparency and disclosure can increase the spillover value g or, alternatively, internalize it through mandated practices and penalties.

At the same time, the model highlights a concrete failure mode for naive mandates. If explanation carries substantial incremental leakage risk in some prompt classes (large Δ_E), then forcing E uniformly can reduce welfare even when E is beneficial on average. This points toward *conditional* governance: require platforms to implement and justify state- or category-dependent refusal protocols, rather than enforcing a single global stance. In our terms, the objective is not to pick E or R once-and-for-all, but to (i) define a defensible robustness index (or vector of indices) that approximates x , (ii) estimate how refusal style affects both short-run misuse and downstream discovery,

and (iii) commit to a switching behavior with documented updates when primitives shift.

We should also be explicit about what our tractable structure leaves out, because those omissions are precisely where the empirical and engineering work concentrates. The one-dimensional state x collapses heterogeneous hazard surfaces into a scalar, yet real systems face category-specific risk and mitigation channels with cross-effects (e.g., a refusal protocol that improves cyber robustness may teach adversarial prompting strategies that increase fraud). Our affine dynamics treat obsolescence as exogenous, whereas adversaries respond strategically to policy, making ω and even the distribution of risky prompts endogenous. We assume perfect observability of x , but deployed platforms operate with noisy, delayed, and sometimes gamed measurements. Each of these extensions can break closed forms and complicate comparative statics, but they do not negate the organizing principle: refusal behavior trades off immediate exposure against the production of safety-relevant information, and that tradeoff is inherently dynamic.

These limitations suggest a concrete research agenda that is both technical and governance-facing. On the technical side, we need methods that (i) estimate learning rates from operational data (how much does a refusal interaction contribute to identifying, reproducing, and patching a vulnerability?), (ii) separate “explanatory content” from “exploit-enabling content” to reduce Δ_E while preserving α_E , and (iii) plan under partial observability and distribution shift, where the sufficient statistic is a belief over robustness rather than x itself. On the governance side, we need verifiable commitments: machine-verifiable logs, evaluation harnesses with versioning and refresh procedures, and audit protocols that can detect when a platform is substituting surface compliance (high refusal rate) for genuine learning (high effective α). In the language of our model, the goal is to prevent a platform from optimizing the wrong objective by hiding in the R region while claiming safety improvements that are not reflected in mitigation kinetics.

A final takeaway is conceptual. Many discussions frame the “helpfulness versus safety” tradeoff as a static Pareto frontier. Our results replace that picture with a dynamic one: helpful-safe refusal can be an *investment* that shifts the frontier outward over time by increasing robustness, but only if explanation is engineered to be diagnostic rather than exploitative and only if institutions reward the resulting spillovers. Conversely, evasiveness can be a form of short-run risk suppression that leaves the underlying vulnerability mass intact, increasing the chance that risk reappears elsewhere or later. In practice, the right question is not “should we refuse more?” but “are our refusals generating the information needed to reduce future harm, and are we measuring that reduction in a way that can be audited and contracted upon?”

If we have succeeded, the value of the model is not that it fully captures deployment reality, but that it isolates a small number of measurable quan-

tities and makes a falsifiable prediction about policy shape. This is the point at which alignment theory can meet governance: once refusal style is treated as a dynamic safety lever with a predictable externality, it becomes possible to design institutions—procurement, audits, transparency, and liability—that push private behavior toward socially efficient learning without naively demanding disclosure that increases exploitation. The remaining work is to build the measurement infrastructure, engineering mitigations, and verification mechanisms that let these abstract primitives be estimated and acted upon in real systems.