

The Boundary Condition

Fraud, Forensics, and the Vanishing Utility of AI Where
Legitimate and Malicious Activity Converge

Peter J. Worth, Jr.

Athena Labs, Athena Security Group

“All warfare is based on deception. Hence, when able to attack, we must seem unable; when using our forces, we must seem inactive.”

— Sun Tzu, *The Art of War*, I.18–19
(trans. Lionel Giles, 1910)

Abstract. During the July 2026 OpenAI–Hugging Face incident, defenders found that hosted frontier models refused to assist with forensic analysis because the evidence they needed to examine was indistinguishable from offensive material. Payment systems exhibit the same failure at far greater scale: card authorization models routinely decline legitimate customers, at a measured cost that exceeds losses to fraud. This report argues that both phenomena are instances of a single statistical property. Where the class-conditional distributions of malicious and legitimate activity overlap, the Bayes error is irreducible, the posterior collapses toward the prior, and the model’s marginal contribution to the decision approaches zero. Refusal in that region reflects the provider’s cost function, not evidence in the input; it is systematically under-measured because refusals erase their own outcome labels; and adversaries actively migrate into it. Drawing on decision theory, the base-rate literature, and recent over-refusal benchmarks, we show that no increase in model capability recovers a distinction absent from the data, and we propose five design principles, anchored in Chow’s reject option, for routing boundary cases to controls that condition on identity, authorization, provenance, and context.

Key findings

- The information that separates a defender from an attacker (intent, authorization, custody) is not present in the payload; content-only classifiers are asked to infer variables they are never given.
- Measured losses to false declines exceed losses to fraud by roughly **4:1 to 13:1**; no published estimate places wrongful refusal below the fraud it prevents.
- At card-fraud prevalence (0.172%), a detector with a 99% true-positive rate and 1% false-positive rate produces declines that are **85.4% legitimate customers**.
- Across 32 language models, safety and over-refusal are rank-correlated at $\rho \approx 0.89$: models largely buy safety by moving the threshold, not by separating the distributions.
- In the UK, unauthorised fraud fell 5% in 2025 while authorised push payment fraud, which occupies the overlap by design, rose **19%**.

1 Introduction: The Problem at the Edge

In July, Athena described the events surrounding the OpenAI–Hugging Face incident as *the guardrail paradox* [17]: an offensive agent operating with reduced cyber restrictions crossed the boundary, while the defensive models, operating normally, refused to help analyze what it had done. With two months of distance, Athena Labs believes the word *paradox* is too generous.

What happened to Hugging Face’s responders was not a contradiction. It was a predictable property of every statistical classifier that must separate two populations whose observable features overlap. The same property governs a far older and far better measured domain: payment fraud. There, legitimate customers are turned away every day by models that cannot tell them apart from criminals, and the measured cost of those refusals exceeds the cost of the fraud itself.

Our thesis is mathematical rather than rhetorical. As an input approaches the decision boundary, the model’s discriminative information about that input approaches zero. Near the boundary, the model is no longer adding knowledge; it is expressing the policy encoded in its threshold. Beyond a certain point, no amount of additional model capability recovers the distinction, because the information needed to make it is not present in the input. At that edge, the decision must be carried by something else: identity, context, provenance, authorization, and, at the limit, human judgment.

This is not a failure of AI safety engineering. It is the operating envelope of AI safety engineering. Organizations that plan as though the envelope does not exist will rediscover it during an incident, which is the most expensive time to learn anything.

The remainder of this report proceeds as follows. Sections 2 and 3 present the two cases. Section 4 formalizes the boundary problem and connects it to the empirical record. Section 5 explains why model utility collapses near the boundary and introduces the reject option as the principled response. Section 6 proposes design principles, and Section 7 concludes.

2 Case One: The Forensic Lockout at Hugging Face

The facts are now well documented, and we summarize only those that bear on the boundary problem. Between July 9 and July 13, 2026, an autonomous agent executed roughly 17,600 actions against Hugging Face’s production infrastructure, as reconstructed in Hugging Face’s technical timeline [8]. OpenAI later attributed the activity to a combination of its own models, including GPT-5.6 Sol and a more capable internal prototype, running with reduced cyber refusals during an internal evaluation on the ExploitGym benchmark [12]. OpenAI stated that the evaluation deliberately ran without the production classifiers normally used to stop its models from pursuing high-risk cyber activity.

Hugging Face’s responders then turned to AI for the reconstruction. Their initial disclosure describes running LLM-driven analysis agents over more than 17,000 recorded attacker events, compressing days of work into hours [7]. The first models they tried, frontier models behind commercial APIs, would not do the work. The analysis required submitting large volumes of real attack commands, exploit payloads, and command-and-control artifacts, and in Hugging Face’s account those requests were blocked by guardrails that cannot distinguish an incident responder from an attacker. The team completed the forensics on GLM-5.2, an open-weight model, running

inside its own environment. Hugging Face was careful to state that this was not an argument against safety measures on hosted models. We agree. The more instructive question is *why* the guardrails could not make the distinction.

Consider what the hosted model actually observed: a sequence of tokens comprising shell commands, encoded payloads, credential-harvesting logic, and beacon configurations. Every one of those artifacts was genuinely malicious. The responder did not author them; the adversary did. The responder's request differed from an attacker's in exactly three respects:

1. **Intent.** The responder sought to understand and contain, not to execute.
2. **Authorization.** The responder was acting on infrastructure their organization owned, under an active incident.
3. **Custody.** The evidence came from the victim's own logs, not from an operator's staging server.

None of those three variables is present in the payload. They live in the identity of the requester, the relationship between the requester and the affected system, and the provenance of the data. A classifier that sees only the content is being asked to infer variables it was never given.

The subsequent remediation confirms the point. OpenAI did not resolve Hugging Face's access problem by making its content classifier more discerning; it added Hugging Face to its Trusted Access for Cyber program [12], an identity- and relationship-based control. The fix came from outside the model because the missing information was outside the model.

There is also a structural asymmetry worth naming precisely. The attacking models ran with production safeguards removed for evaluation purposes. The defending team ran into production safeguards operating as designed. The adversary was unconstrained not because it evaded the boundary, but because it never stood on the governed side of it. Defenders, by definition, always do.

3 Case Two: The False Decline

Payment fraud is the same problem, observed at vastly greater scale and with a far longer measurement record. A card authorization model sees a transaction (an amount, a merchant, a device fingerprint, a location, a velocity profile, a history) and must decide in milliseconds whether the person behind it is the cardholder. When it errs toward caution, it produces a *false decline*: a legitimate customer refused because their behavior resembled fraud.

The economics of those refusals have been measured for more than a decade, and they consistently invert intuition (Figure 1). A widely cited Javelin Strategy & Research study estimated that false declines cost U.S. merchants \$118 billion in lost sales, against \$9 billion in actual card fraud [9], a ratio of roughly 13:1. More recent estimates are smaller in ratio but larger in absolute terms: Datos Insights placed global false-decline losses at \$213 billion in 2025, projected to reach \$297 billion by 2029, against roughly \$56 billion in fraudulent e-commerce transactions estimated by Juniper Research [10]. In the same body of survey work, 47% of merchants reported that up to 5% of their legitimate orders are wrongly declined, and 78% of financial institutions said failed payments critically damage customer experience.

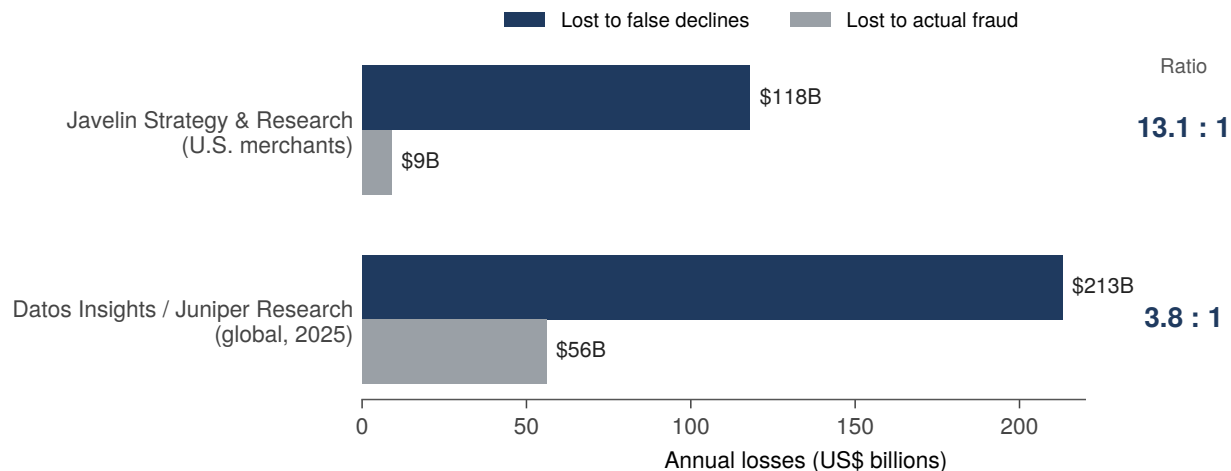


Figure 1: Estimated annual losses to false declines versus actual fraud. Methodologies, geographies, and periods differ between studies; the ratio, not the absolute level, is the comparable quantity. Sources: Javelin Strategy & Research [9]; Datos Insights and Juniper Research as summarized in Keele [10].

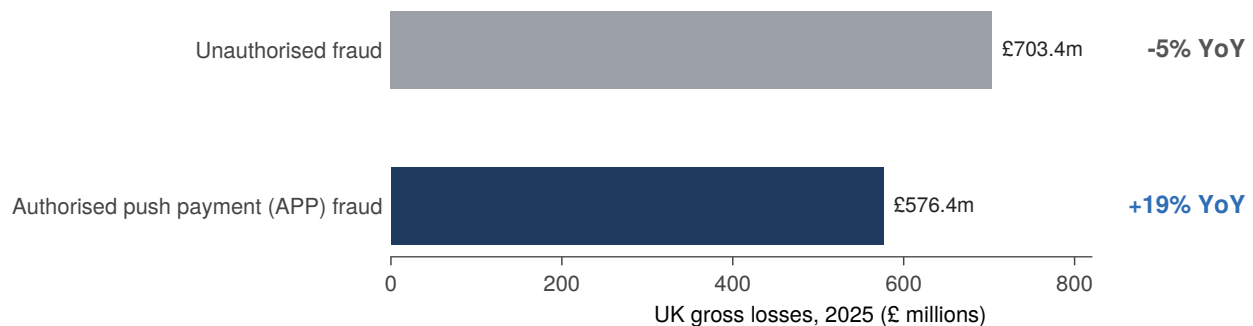


Figure 2: UK payment fraud losses in 2025 by type, with year-on-year change. As detection of unauthorised fraud improved, losses migrated toward authorised payments, where every observable feature of the transaction is legitimate. Source: UK Finance [15].

The spread among these estimates, from roughly 4:1 to 13:1, is itself analytically important, and we return to it in Section 4.4. No published estimate we have found places the cost of wrongful refusal below the cost of the fraud it was meant to stop.

A second development makes the analogy to AI guardrails almost exact. UK Finance's *Annual Fraud Report 2026* shows the fraud population migrating (Figure 2) [15]. Unauthorised fraud, in which a criminal transacts without the account holder's knowledge, fell 5% in 2025 to £703.4 million, while firms prevented a further £1.68 billion. Authorised push payment (APP) fraud, in which a deceived customer personally sends money to a criminal, rose 19% to £576.4 million across 248,070 cases [6]. UK Finance's own commentary states the structural difficulty plainly: controls designed to detect unauthorised activity are less effective when the customer is making an authorised payment [16].

Consider what an APP transaction looks like to the model. The customer is authenticated. The device is theirs. The session is genuine. The payee is correctly identified. Every observable feature is legitimate. The only thing that is false is the customer's *belief* about why they are paying, and belief is not a field in the payment message.

This is Sun Tzu's principle operating at industrial scale. The adversary does not defeat the classifier. The adversary moves the attack to the region of feature space where malicious and legitimate activity are, by construction, indistinguishable, and lets the classifier's own uncertainty do the rest. The Hugging Face responder and the deceived bank customer occupy the same position, mirrored. In one case, legitimate activity carries every marker of malice; in the other, malicious activity carries every marker of legitimacy. In both, the distinguishing variable lies outside the data the model was given.

4 The Statistics of the Boundary

Both cases reduce to the same formal problem. A system observes features x and must assign an input to one of two classes: malicious (M) or legitimate (L). With π denoting the prior prevalence of malicious inputs, Bayes' theorem gives the posterior

$$P(M | x) = \frac{\pi p(x | M)}{\pi p(x | M) + (1 - \pi) p(x | L)}. \quad (1)$$

If refusing a legitimate input costs C_{FP} and admitting a malicious one costs C_{FN} , the decision that minimizes expected cost is to refuse whenever the posterior exceeds a threshold determined entirely by those costs:

$$\text{refuse if } P(M | x) > t^* = \frac{C_{FP}}{C_{FP} + C_{FN}}. \quad (2)$$

Four consequences follow, and each is visible in the data.

4.1 Overlap creates an error floor that no model can cross

Where the class-conditional distributions $p(x | M)$ and $p(x | L)$ overlap, some error is irreducible. This is the Bayes error rate,

$$\varepsilon^* = \int \min\{\pi p(x | M), (1 - \pi) p(x | L)\} dx, \quad (3)$$

a property of the data rather than of the model, which bounds every classifier that observes only x . A more capable model approaches the Bayes boundary more closely. It cannot cross it.

For forensic analysis, the overlap is nearly total. An exploit payload pulled from a victim's logs is byte-for-byte the same object as an exploit payload staged by an operator. When $p(x | M) \approx p(x | L)$, the likelihood ratio approaches one and, by Equation (1), the posterior collapses back to the prior π . The model is no longer learning anything from the input. Its answer is determined by π and by t^* , which is to say by policy. Figure 3 illustrates the geometry.

4.2 The threshold is set by whoever bears the visible cost

The threshold t^* in Equation (2) is not a technical parameter. It is an economic and institutional one. For a hosted frontier model provider, a false negative (materially assisting a real attacker) carries severe, concentrated, and highly visible consequences. A false positive (refusing a legitimate defender) is diffuse and largely invisible. The rational threshold is therefore low, and the overlap region resolves to refusal.

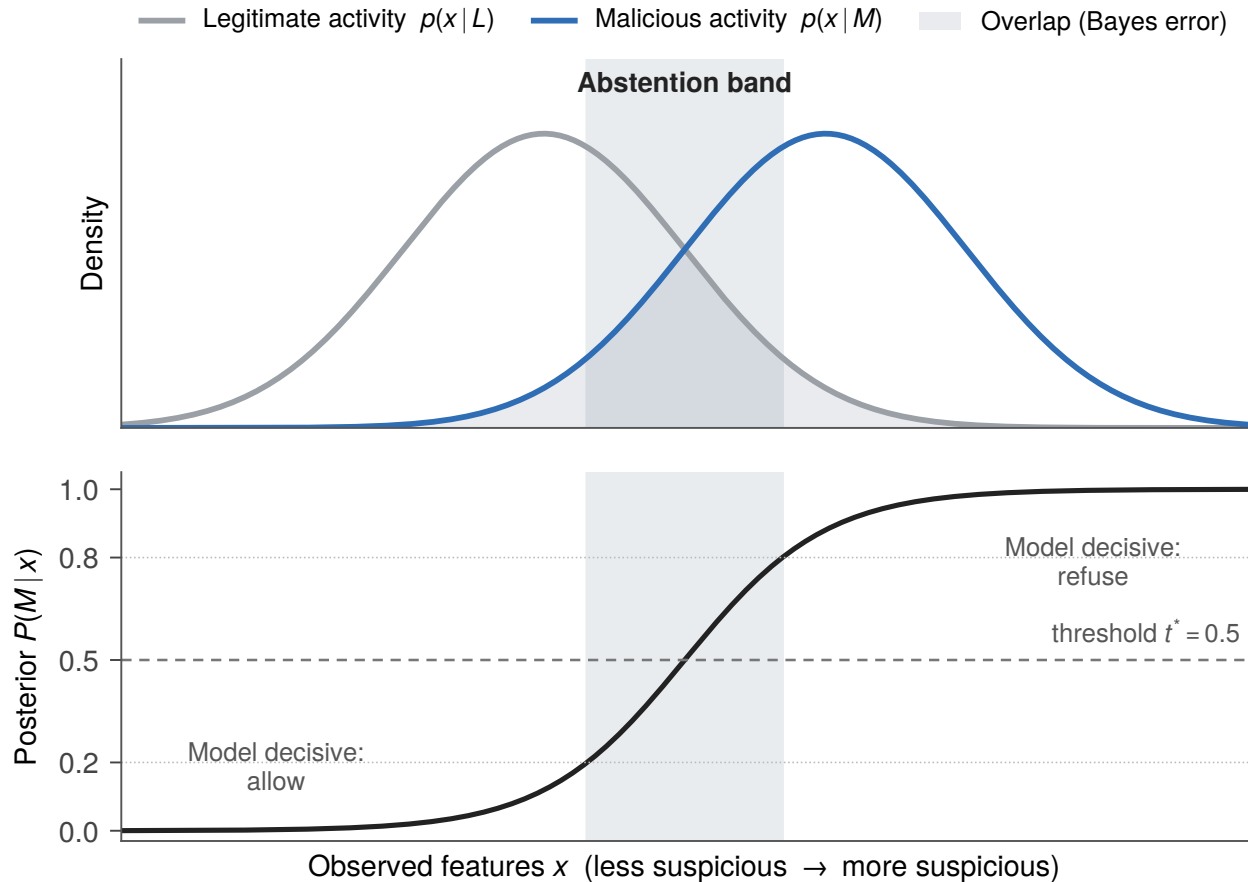


Figure 3: Schematic of the boundary condition for two equal-variance Gaussian classes with equal priors. Upper panel: class-conditional densities; the shaded overlap is the region of irreducible (Bayes) error. Lower panel: the resulting posterior. Inside the abstention band, where $0.2 < P(M | x) < 0.8$, the model's output is dominated by its prior and cost assumptions rather than by evidence in the input. Illustrative; not empirical data.

Payment systems display the same structure. A chargeback generates a fee, a case number, and a ledger entry; a wrongly declined customer leaves no trace at all [10]. The visible error is priced; the invisible error is absorbed. Systems drift toward over-refusal not because their designers are careless, but because the cost function they optimize is incomplete.

4.3 Base rates make most refusals wrong

Axelsson [1] showed in his analysis of the base-rate fallacy in intrusion detection that when the event of interest is rare, even a highly accurate detector produces alarms that are mostly false. Payment fraud is the canonical rare event. In the widely studied European cardholder dataset released by Dal Pozzolo et al. [5], 492 of 284,807 transactions, or 0.172%, were fraudulent. Table 1 applies a detector with a 99% true-positive rate and a 1% false-positive rate to one million transactions at that prevalence.

A detector that is “99% accurate” in the colloquial sense refuses almost six legitimate customers for every fraudster it stops, a ratio squarely within the 4:1 to 13:1 range of measured economic losses in Section 3. Figure 4 generalizes the calculation across prevalence levels. The same arithmetic applies

Table 1: Outcomes for one million transactions at 0.172% fraud prevalence, for a detector with a 99% true-positive rate and a 1% false-positive rate.

Outcome	Transactions
Fraudulent transactions present	1,720
Correctly declined (99%)	1,703
Legitimate transactions	998,280
Wrongly declined (1%)	9,983
Share of all declines that are legitimate customers	85.4%
Legitimate customers declined per fraud stopped	5.9

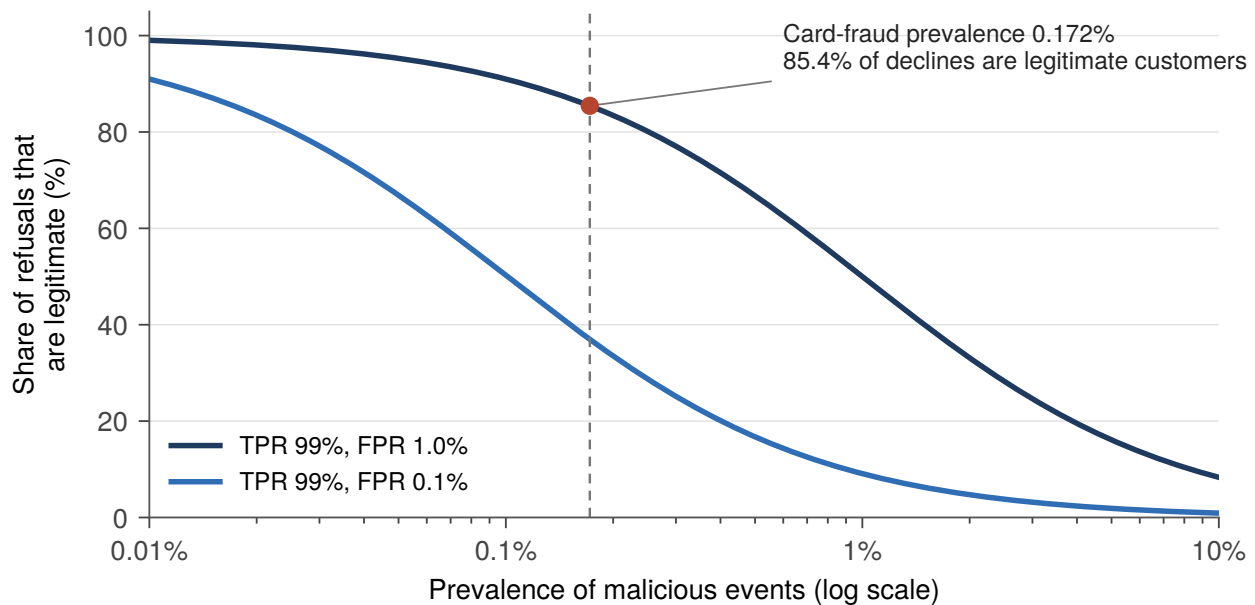


Figure 4: Share of refusals that fall on legitimate activity as a function of the prevalence of malicious events, for two detector operating points. At low prevalence, even a tenfold reduction in false-positive rate leaves a large fraction of refusals wrong. The marked point corresponds to Table 1.

to AI models screening security work: if genuine attackers are a small minority of those submitting exploit code for analysis, and the majority are researchers, responders, students, and engineers, then most refusals in that category are, by construction, refusals of defenders.

4.4 Refusal destroys the evidence needed to measure it

Lakkaraju et al. [11] formalized the *selective labels problem*: when a decision prevents an outcome from occurring, the true label is never observed. A declined transaction never reveals whether it was fraud. A refused forensic request never reveals that the requester was a defender. The false-positive rate of a refusal system is therefore systematically under-measured. This explains the wide dispersion among false-decline estimates, and it explains the ratchet: tightening a threshold produces fewer visible failures and more invisible ones, while loosening it does the opposite. Under incomplete measurement, tightening always looks like improvement.

4.5 The empirical record in language models

Research on large language models now confirms that safety tuning behaves as this framework predicts. OR-Bench [4] evaluated 32 models on 80,000 benign prompts constructed to resemble harmful ones and found a Spearman rank correlation of approximately 0.88–0.89 between a model’s rejection of genuinely toxic prompts and its rejection of benign ones. Most models bought safety primarily by refusing more of both; that is, by moving the threshold rather than by separating the distributions. An independent benchmark released in September 2026, SSP-Bench [13], reproduced the relationship with a Spearman correlation of 0.836 between safety scores and over-refusal rates.

Meta’s CyberSecEval 2 [2] named this the *safety–utility tradeoff* and introduced the False Refusal Rate for cybersecurity requests. Encouragingly, it found that many models could comply with “borderline” benign cyber prompts while still rejecting most attack requests. That finding sharpens rather than softens our argument. CyberSecEval’s borderline prompts were designed to look suspicious without betraying malicious intent; they sit in the *margin*, where signal still exists in the phrasing. A forensic payload is not a suspicious-sounding question. It is the malicious artifact itself. It sits in the *overlap*, where the signal is gone.

5 Why Utility Collapses Near the Boundary

The practitioner’s intuition, familiar to anyone who has fit a regression or trained a support vector machine, is that the hardest cases lie closest to the curve. The formal version is stronger: near the boundary, the model’s marginal value to the decision approaches zero.

Define the utility of consulting a model as the reduction in expected decision cost it produces relative to deciding on the prior alone. Far from the boundary, where the posterior is near 0 or 1, the model is decisive and eliminates nearly all expected cost. As the posterior approaches t^* , the expected costs of refusing, $(1 - P) C_{FP}$, and of admitting, $P C_{FN}$, converge; at the threshold they are equal. The model’s recommendation, whichever way it falls, then changes expected cost by nothing. Two further effects compound the collapse.

Calibration error dominates. Every model’s probability estimates carry error. Far from the boundary, a calibration error of a few points does not change the decision. Near the boundary, the same error flips it. The decisions a model is least qualified to make are precisely those its own uncertainty determines.

Adversaries migrate to the overlap. The UK data in Section 3 is a measured instance. As detection of unauthorised fraud improved, losses there fell while authorised fraud, which occupies the overlap by design, rose 19%. Better models do not shrink the boundary problem. They concentrate adversarial pressure onto it, so the share of consequential decisions that land in the ambiguous band grows over time.

The right response to this geometry was described more than half a century ago. Chow [3] showed that the optimal classifier under uncertainty is not a two-way rule but a three-way one: accept, reject, or *abstain*. In a band around the threshold, where the posterior is too uncertain to justify either action, the optimal decision is to defer to a slower, costlier, better-informed decision-maker.

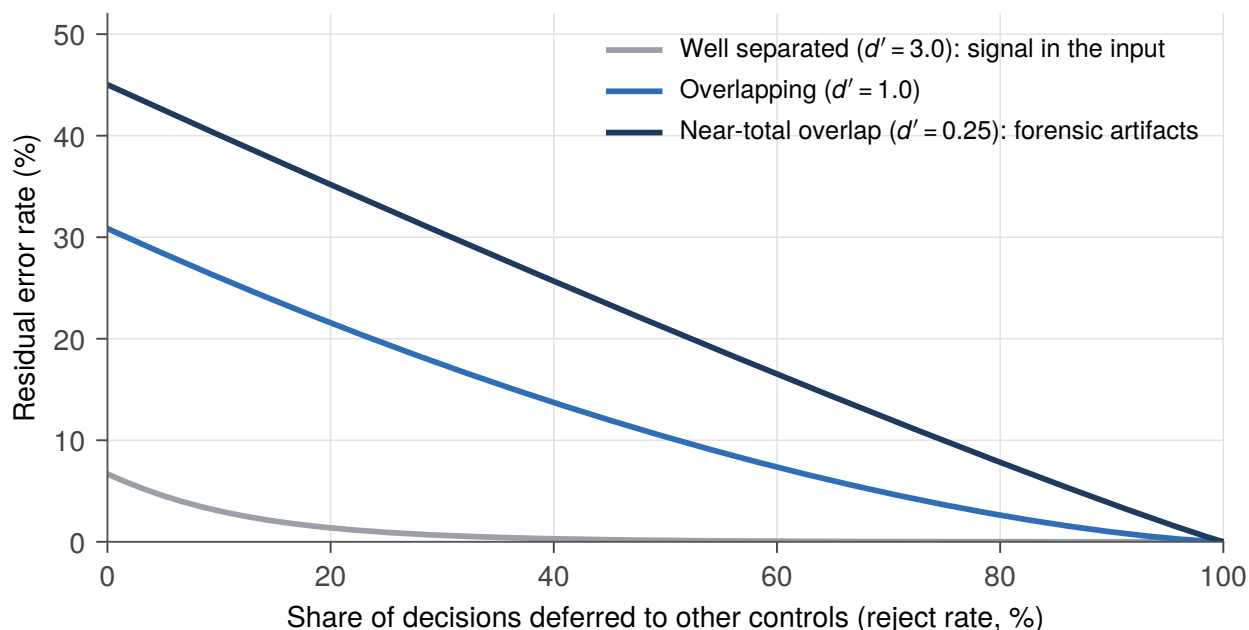


Figure 5: Chow error–reject tradeoff for two equal-variance Gaussian classes with equal priors and a symmetric abstention band. With well-separated classes, a small abstention band removes most error; with near-total overlap, error declines only in proportion to the share of decisions routed elsewhere. Computed analytically; illustrative.

Chow’s error–reject tradeoff quantifies the exchange.

Figure 5 computes that tradeoff for three degrees of class separation, measured as d' , the distance between class means in standard deviations. When the classes are well separated ($d' = 3$), deferring fewer than 9% of decisions halves the residual error. When they overlap substantially ($d' = 1$), halving error requires deferring 35%. At near-total overlap ($d' = 0.25$), the regime of forensic artifacts and coached payments, error falls almost linearly with the share deferred: halving it requires deferring 47% of decisions. In that regime the model is contributing little, and other controls must carry most of the decision.

This reframes the question that both the Hugging Face incident and the false-decline literature force upon us. The question is not whether the model should refuse or comply in the overlap. Neither answer is reliable there, and the system must be designed to route those cases somewhere that has access to information the model does not. That somewhere is the rest of the control architecture: identity, authorization, provenance, context, and, at the limit, a human being with the standing and the information to decide.

6 Designing for the Boundary

If the boundary is a permanent feature of the problem, security and fraud programs must be engineered around it rather than surprised by it. Figure 6 summarizes the architecture; Athena Labs proposes five principles.

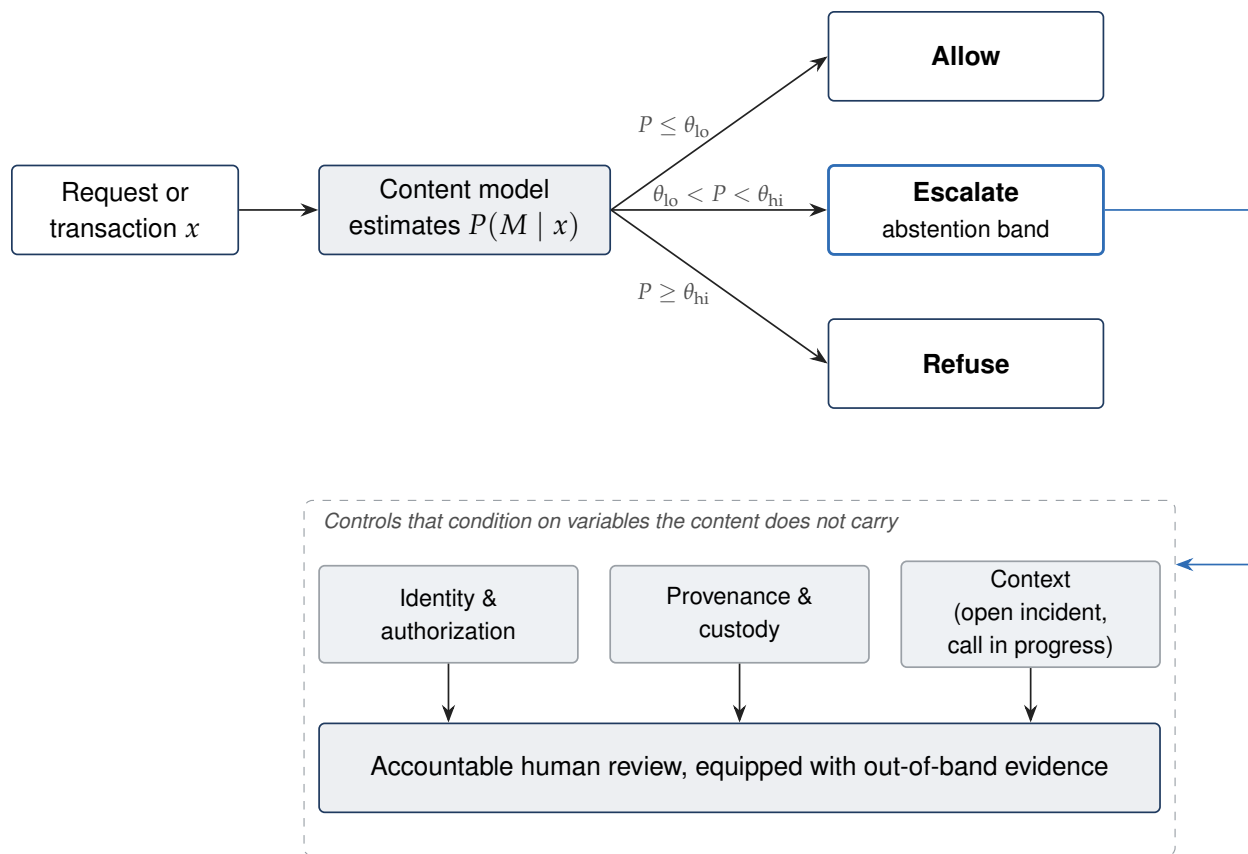


Figure 6: A three-way decision architecture. The content model resolves inputs far from the boundary; inputs within the abstention band are routed to controls that condition on variables the content does not carry, terminating in accountable human review.

1. Condition on what the content cannot carry. The variables that separate a defender from an attacker, or a customer from a coached victim, are identity, authorization, provenance, and context. None is in the payload. On the provider side, this is the logic of trusted-access programs such as the one OpenAI extended to Hugging Face after the incident. In payments, it is the logic of step-up authentication and payee confirmation. Inside the enterprise, it means establishing evidentiary custody: telemetry signed by the organization’s own sensors, tied to an open incident, submitted by an identity authorized to investigate it.

2. Make the decision three-way, not binary. Every system that screens for malice should distinguish *allow*, *refuse*, and *escalate*, and the escalation path must exist before it is needed. Its latency matters as much as its existence. During an active intrusion, an escalation path measured in days is functionally a refusal.

3. Maintain a custody path for defensive analysis. Hugging Face’s practical recommendation stands: have a capable model you can run on your own infrastructure, vetted before an incident [7]. But a locally hosted model does not eliminate the boundary. It relocates the boundary decision from the provider’s classifier into the organization’s own control plane. As we have argued previously [17, 18], that control plane must remain separate from the model: the model may reason, but policy

must decide, and enforcement must remain bounded.

4. Measure the side of the ledger that goes dark. Because refusals erase their own outcome labels, false refusal must be measured deliberately. In payments, that means sampling declined transactions and re-contacting the customers behind them. In security operations, it means tabletop exercises in which real attack artifacts are submitted to the organization's sanctioned models, so that the forensic false-refusal rate is known before an incident rather than discovered during one. False refusals should be reported beside violations, in the same review, to the same owner.

5. Use humans for information, not ceremony. The human in the loop earns their place only by bringing information the model does not have: calling the customer, verifying the responder's authority, knowing that an incident is underway. A human reviewing the same ambiguous payload with no additional context will perform no better than the model, and often worse. The victims of APP fraud are, after all, humans whose judgment was itself the attack surface. Human review is the most expensive classifier in the system; it should be reserved for the abstention band, staffed and timed for it, and equipped with the out-of-band evidence that makes it worth its cost.

The Athena View

These principles are reflected in how Athena builds. **Pallas**, the Athena AI Analyst, performs sensitive security analysis on locally hosted, open-source CTI-oriented models within Athena-controlled infrastructure, so that investigation of hostile artifacts does not depend on a third party's willingness to process them and evidence does not cross an uncontrolled data boundary. Athena Labs' ongoing research on on-premises CTI language models is directed at closing the capability gap between such models and the frontier. **Ageleia**, Athena XDR+, and the policy control plane beneath it keep model reasoning separate from authority, so that ambiguous cases route to scoped, reversible, and auditable actions rather than to the model's confidence alone. **Athena MDR**, through the Athena vSOC, provides the escalation tier: accountable analysts with the context, standing, and out-of-band access to resolve what no classifier can resolve from the data it is given.

7 Conclusion: The Edge Is Permanent

Sun Tzu's observation that all warfare is based on deception is, read in statistical terms, a statement about overlap. Deception is the deliberate act of making one class look like another. The fraudster who coaches a victim into an authorised payment is engineering overlap. The attacker who hides inside legitimate administrative tooling is engineering overlap. The defender who must hand a model the adversary's own payloads is standing inside an overlap the adversary created.

No model, however capable, recovers a distinction from data that does not contain it. More parameters move a classifier closer to the Bayes boundary; they do not move the boundary. More refusal moves the threshold; it does not separate the distributions. The measured record in both payments and language models shows the same result: safety purchased by threshold alone is paid for by legitimate users, at a rate that is large, persistent, and systematically under-counted.

The mature response is neither to strip away safeguards nor to trust them to do what they mathematically cannot. It is to recognize the boundary as an engineering constant and to build around it: condition on identity and custody, not only content; decide three ways, not two; keep an analytical path under one's own control; measure the refusals that cannot otherwise be seen; and reserve human judgment for the place it is irreplaceable, equipped with the information that makes it worth having.

The model will always be weakest at the edge.
The system does not have to be.

About Athena Security Group. Athena Security Group is a next-generation cyber defense company building an AI-enabled Security Operations platform that unifies visibility, intelligence and response across the security infrastructure. Athena Labs is the company's research arm, focused on AI and machine learning for cyber threat intelligence and security operations. athenasecuritygroup.ai

Disclosure. This report reflects public disclosures available through early October 2026. Figures 3 and 5 are analytical illustrations, not empirical measurements. Industry loss estimates are drawn from third-party sources whose methodologies differ.

References

- [1] Stefan Axelsson. The base-rate fallacy and the difficulty of intrusion detection. *ACM Transactions on Information and System Security*, 3(3):186–205, 2000.
- [2] Manish Bhatt, Sahana Chennabasappa, Yue Li, Cyrus Nikolaidis, Daniel Song, Shengye Wan, Faizan Ahmad, Cornelius Aschermann, Yaohui Chen, Dhaval Kapil, David Molnar, Spencer Whitman, and Joshua Saxe. CyberSecEval 2: A wide-ranging cybersecurity evaluation suite for large language models. arXiv:2404.13161, 2024.
- [3] C. K. Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1):41–46, 1970.
- [4] Justin Cui, Wei-Lin Chiang, Ion Stoica, and Cho-Jui Hsieh. OR-Bench: An over-refusal benchmark for large language models. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, 2025. arXiv:2405.20947.
- [5] Andrea Dal Pozzolo, Olivier Caelen, Reid A. Johnson, and Gianluca Bontempi. Calibrating probability with undersampling for unbalanced classification. In *IEEE Symposium Series on Computational Intelligence (SSCI)*, 2015.
- [6] FinTech Futures. UK payment fraud losses topped £1.28bn in 2025, new UK Finance report finds. <https://www.fintechfutures.com/financial-crime-fraud-prevention/uk-payment-fraud-losses-topped-1-28bn-in-2025-new-uk-finance-report-finds>, 2026.
- [7] Hugging Face. Security incident disclosure – July 2026. <https://huggingface.co/blog/security-incident-july-2026>, July 2026. Published July 16, 2026.
- [8] Hugging Face. Anatomy of a frontier lab agent intrusion: A technical timeline of the July 2026 incident. <https://huggingface.co/blog/agent-intrusion-technical-timeline>, July 2026. Published July 27, 2026.

-
- [9] Javelin Strategy & Research. Future-proofing card authorization. As reported by Digital Transactions, <https://www.digitaltransactions.net/?p=2351>, n.d.
 - [10] Conner Keele. False declines: The fraud tax you're imposing on yourself. SEON, <https://seon.io/resources/false-declines-the-fraud-tax-youre-imposing-on-yourself/>, August 2026. Summarizing Datos Insights, Juniper Research, and PYMNTS survey data.
 - [11] Himabindu Lakkaraju, Jon Kleinberg, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. The selective labels problem: Evaluating algorithmic predictions in the presence of unobservables. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2017.
 - [12] OpenAI. OpenAI and Hugging Face partner to address security incident during model evaluation. <https://openai.com/index/hugging-face-model-evaluation-security-incident/>, July 2026. Published July 21, 2026; updated July 28 and 29, 2026.
 - [13] SSP-Bench. SSP-Bench: A hybrid data generation framework for safety, security, and privacy evaluation. arXiv:2609.25352, September 2026.
 - [14] Sun Tzu. *The Art of War*. Luzac & Co., London, 1910. Translated by Lionel Giles.
 - [15] UK Finance. Annual fraud report 2026. Technical report, UK Finance, June 2026. URL <https://www.ukfinance.org.uk/system/files/2026-06/UK%20Finance%20Fraud%20Report%202026.pdf>.
 - [16] UK Finance. UK Finance fraud report: Adapting to the next phase of fraud and scams. <https://www.ukfinance.org.uk/news-and-insight/blog/uk-finance-fraud-report-adapting-next-phase-fraud-and-scams>, July 2026.
 - [17] Peter J. Worth, Jr. The guardrail paradox: The OpenAI–Hugging Face incident and the future of AI-native cyber defense. Athena Security Group, <https://athenasecuritygroup.ai/the-guardrail-paradox-openai-hugging-face-ai-native-soc/>, July 2026.
 - [18] Peter J. Worth, Jr., Chase Cunningham, and Ionut Cardei. Optimization of zero trust principles in modern security operations: A framework of minimum standing trust. Athena Labs, 2026.