How to Train Your AI When Humans Matter

Kevin A. Bryan and Joshua S. Gans

March 2026

Abstract

AI predicts; its prediction is used by humans in decisions. The value of AI therefore depends on how humans extend, verify, and act on that prediction. We model AI as part of a composite experiment where agents can verify predictions at cost, delegate to AI, or avoid the model altogether. We derive the optimal coverage–conditional accuracy tradeoff for training, showing that maximizing unconditional accuracy is generally suboptimal. The optimum is discontinuous where users switch between autonomous and verified AI regimes. Heterogeneous users disagree on the ideal model. Because the optimal model maps directly from the economic environment - downside risk, verification cost, adversarial pressure, task complementarities - knowledge of the economic environment can inform about the nature of optimal training ex-ante.

JEL: D81, D82, O33.

Keywords: artificial intelligence, selective prediction, verification, decision theory, reject option, training frontiers.

1 Introduction

What makes an AI - a prediction machine - economically valuable? The simple but wrong answer is "a model which predicts accurately as often as possible." This is wrong because the predictions made by an AI are used as inputs into an agent's decision function, alongside human estimation, information from other models, sequential verification, the ability to abstain from acting, and so on.

Many problems fit this setting. A surgeon uses a medical diagnostic tool to predict how to proceed in a complicated surgery, at the cost of a small delay. The prediction is perhaps inaccurate. The doctor may attempt to verify the prediction at cost, or simply act on the prediction. An error harms the patient. What does a good diagnostic AI look like? Or consider three ways to drive a car. A black-box self-driving system makes the right decision 95% of the time but provides little guidance about which 5% of situations produce failures. A cruise-control system drives perfectly on highways but cannot drive in cities at all. A human driver can drive in both environments but at a high effort cost. Which car provides more economic value? The important thing to realize in these situations is that the model's prediction is an input into a decision function, not the action itself.

This problem is well-known in the machine learning literature, where ideas like selective classification (Chow, 1970) and learning-to-defer (Madras et al., 2018) attempt to maximize value from a model that can say "I don't know" when its prediction is weakly informative, or can defer to an agent to make a final decision. These methods generally try to *infer* the relative value of the human decision-maker jointly with the model. We consider this problem from the other side: take an agent, or set of agents, with a known decision function, and then investigate what model parameters maximize utility. That is, we model AI-human interaction as a problem of composite experiments in statistical decision theory. The optimal AI will depend on the human's loss function, their cost of verification, and any externalities that might arise in strategy settings.

Our primary results are threefold. What do agents do in response to a prediction? How should we then train the model? Which agents benefit the most? These questions map to three distinct design margins. *Use design* determines whether a user should trust, verify, or ignore a given prediction. *Training design* determines which point on the accuracy–coverage frontier the developer should target. *Reporting design* determines whether the model emits only a pooled recommendation or also exposes case-specific reliability information. The paper proceeds in that order: it first characterizes use, then training under coarse outputs, then the consequences of heterogeneous users, and finally the way richer posterior reporting changes what should be trained and how outputs should be presented.

What makes a good AI for the agent? We show that only three canonical strategies are ever optimal: abstain from using the AI, trust its recommendation, or verify its prediction before acting. Conditional on receiving a prediction, trusting without verification is optimal only when verification is expensive relative to the expected loss from erroneous actions *and* acting on the prediction beats reverting to the baseline. Even very accurate AI may not be used when the loss from errors is severe and the cost of verification is high. This decision between trusting, verifying, or abstaining from a given AI depends on parameters of the decision function. We investigate exactly how it varies in adversarial, competitive, O-ring, and other settings.

What should the AI model's training target be? We describe model training as choosing a point on a concave accuracy-coverage frontier. Economic value is strictly increasing in conditional correctness holding unconditional accuracy fixed. This means maximizing unconditional accuracy is generically suboptimal. The relevant training levers are refusal behavior, thresholding, and other design choices that move the model along this frontier. The interior optimum lies at a tangency between the frontier and an iso-payoff line - but the slope of that line differs across the trust and verify regimes. In the trust regime, the tangency condition is Chow's selective-classification threshold, applied at the margin of the training frontier rather than the margin of deployment. In the verify regime, coverage should expand until the marginal accuracy gain equals the verification cost per unit of potential gain. The optimally trained model can be *discontinuous* in model and agent parameters due to regime switching, meaning slightly different environments may mean vastly different tradeoffs between coverage and conditional accuracy. For instance, a medical AI whose users can cheaply run confirmatory tests should be trained to high coverage; the same AI, deployed where verification is infeasible, should be trained to high selectivity. At the point where confirmatory tests just become too expensive, the optimal coverage jumps from high to low.

Do users agree which model is best? Not in general. Without perfectly calibrated scores, there is no universal model ranking in the Blackwell sense. Agents with different downside risks or verification costs will disagree about which model they prefer, and this disagreement is not resolved by maximizing unconditional accuracy. When a monopolist trains a single model for a heterogeneous population, the model is trained to the preferences of the marginal paying customer rather than the population average, which can mean too much coverage (and too many errors) for high-risk users, or too much selectivity for low-risk ones. Moving to calibrated score reporting changes the design problem: output usefulness can now come from exposing informative uncertainty, not only from choosing a single common coverage level. For Blackwell-comparable posterior experiments, that restores unanimous

ranking, but it does not resolve disagreement about the training frontier itself.

We say "AI" here rather than "statistical model" even though the latter are of course also inputs into decision functions. Our reduced form is intentionally agnostic: it can represent a large language model forced to answer every query, a selective classifier that abstains on uncertain cases, or a traditional statistical procedure that only applies when identifying assumptions hold. The interesting property of AI, however, is the high cost of training individual models. This means that heterogeneous agents draw on homogeneous models, hence the nature of how those models are trained is welfare-relevant. It is also worth noting that the "regime-switching" discontinuity in optimal model training and the heterogeneity results will hold with our structure extended to include, for example, human augmentation of models, or combinations of information from different sources, or many other human-plus-AI interactions.

We will begin with the coarsest setting, in which the user sees only a pooled prediction, because that isolates the basic trust/verify/abstain problem and makes clear why training must trade off coverage against conditional correctness. We then add heterogeneity to show why one common model can mis-target users, and then add state-specific posteriors to show when better reporting design lets users perform the selection task for themselves. Monopoly and the applications then ask how market structure and surrounding institutions distort those same design margins. Table 1 summarizes this progression.

Related work in economics studies prediction as an input into decision making and the allocation of decision authority between humans and algorithms. Agrawal et al. (2018) emphasizes that many economic changes from AI can be understood as a fall in the cost of prediction, while Athey et al. (2020) studies how a principal allocates authority between a human agent and an AI when human effort is endogenous. Agrawal et al. (2025); Catalini et al. (2026); Agrawal et al. (2026) have similarly examined complementary skills and human-judgment elements that affect decision-making with AI. A number of economists have studied the problem of how to compare an AI to a human decision maker, especially when human decisions generated the data, though this is outside the pure setting of composite experiments (e.g., Kleinberg et al. (2017)). Where agents are biased or misaligned, Agarwal et al. (2025) studies the optimal design of human-AI collaboration using a sufficient-statistic approach that accommodates biased human beliefs (e.g., overconfidence in own-signals) and effort crowd-out. Empirically, recent evidence finds large productivity gains from generative AI in some tasks but not others, and more generally points to substantial heterogeneity across tasks, workers, and human–AI deployment settings (Brynjolfsson et al., 2025; Noy and Zhang, 2023; Vaccaro et al., 2024). This heterogeneity is consistent with the view that the value of AI depends on the surrounding decision problem, including the ease of oversight and the

Table 1: Roadmap of the paper

Part of paper	**Main result(s)**	**What this part does**
Section 2: decision problem	Lemma 1; Propositions 1–2	Sets up the baseline environment. Shows that only three canonical responses to AI are ever optimal: ignore it, trust it, or verify it. Characterizes when trust beats verification, when those queried actions also beat the baseline, and when the model is not used at all.
Section 2.5 and Section 2.6	Definition 1; Lemma 2	Connects the model to selective classification and introduces the training-possibilities frontier $A(q)$. Shows how expanding coverage lowers conditional correctness and can induce a single switch between trust and verify regimes.
Section 3 and Section 3.1	Proposition 3; Corollary 1; Proposition 4	Characterizes optimal training on the frontier. The key result is that maximizing unconditional accuracy is generally not optimal: the best model balances coverage and conditional correctness according to the surrounding decision problem, with different tangency conditions in the trust and verify regimes.
Section 3.2	Proposition 5; Corollary 2; Remark 1	Introduces heterogeneity across users. Shows that different agents generally prefer different points on the frontier, so there is no environment-invariant ranking of models in the reduced-form setting, and a common model creates welfare losses away from each agent's ideal coverage.
Section 3.3	Propositions 6–10; Corollary 3	Enriches the model by allowing observable state-specific posteriors. This lets users condition trust, verification, and abstention on confidence scores, restores unanimous ranking for Blackwell-comparable posterior experiments, and clarifies when posterior informativeness matters more than scalar accuracy.
Section 3.4	Proposition 11	Studies monopoly training. Shows that a monopolist trains for the marginal paying user rather than the social optimum, generating distortions in coverage and selectivity when users differ in risk and verification cost.
Section 4	Propositions 12–17	Applies the baseline logic to a series of environments: adversarial manipulation, imperfect verification, O-ring tasks, committees, payoff externalities across users, and limited liability. Each application works by changing either the feasible frontier or the relevant trust/verify tradeoff.

consequences of error, though the cited studies do not separately identify those channels.

Our results also relate to a well-known literature in machine learning on selective classification and "learning to defer" (L2D). Selective classification, due to Chow (1970), considers a task where an algorithm must assign items to one of a set of labels. The model generates a posterior probability that the classification is correct, and may abstain from classifying if, given the cost of errors versus abstention, that probability is below a threshold. This has been extended to L2D (Herbei and Wegkamp, 2006; Bartlett and Wegkamp, 2008; Cortes et al., 2016; Madras et al., 2018; Mozannar and Sontag, 2020), where the model is trained to maximize the joint accuracy of an algorithm that can either make a prediction, or abstain and pass to an imperfect human expert. The key difference is that these models are largely concerned with how to algorithmically *learn* when to abstain or defer on the basis of observed outcomes. Our model starts from the economic frame that agents have preferences and technologies - in this case, a composite experiment - and asks how the agent will use a given prediction, and hence how to design the model to maximize economic value.[1]

2 The Decision Problem

We begin with the coarsest environment, where the model's output is summarized only by whether it predicts and by how often it is correct when it does. This is the case in which the user cannot sort predictions by quality after the fact, so almost all of the selection problem must ultimately be solved by how the model is trained.

A decision maker faces a single decision under uncertainty. The true state is $\omega \in [0,1]$. An action is $x \in [0,1]$ or abstention $\varnothing$. Taking $x = \omega$ yields payoff $H > 0$. Taking $x \neq \omega$ yields a random payoff L drawn from a distribution F supported on $(-\infty, L_{\max}]$ with $L_{\max} < H$. Let $\bar{L} := \mathbb{E}[L]$. Abstention yields payoff 0.

Two costly human actions are available:

(i) **Solving.** At cost $c_S > 0$, the decision maker learns ω and chooses $x = \omega$, yielding $H - c_S$.

(ii) **Verification.** Given a proposed state $\hat{\omega}$, the decision maker pays $c_V \geq 0$ to learn whether $\hat{\omega} = \omega$ (a binary signal). Verification is perfect.[2]

Define the *baseline* value

$$B := \max\{H - c_S,\ 0\}. \tag{1}$$

[1] A recent paper by Kalai et al. (2025) highlights persistent accuracy issues surrounding large language models and generative AI and argues for confidence-threshold rules that determine when those models should refrain from producing an output.

[2] We will consider imperfect verification in Section 4.

The AI model is queried at cost $c \geq 0$. The model is a triple $m = (q, p, c)$, where $q = \Pr(\hat{\omega} \neq \varnothing)$ is coverage (the probability that the model predicts at all) and $p = \Pr(\hat{\omega} = \omega \mid \hat{\omega} \neq \varnothing)$ is conditional correctness (in the machine learning literature, $1 - p$ in this setting is often called *selective risk*). Unconditional accuracy is $a := qp$. The key reduced-form assumption is that conditional on the event that the model predicts, the realized recommendation carries no additional case-specific information about whether it is correct.

Assumption 1 (Standing assumptions). Throughout the paper: (i) $\bar{L} < B$, so an incorrect action is strictly worse in expectation than the baseline; (ii) whenever we evaluate a queried reduced-form model, $q > 0$. The boundary case $q = 0$ is nevertheless feasible when we later study the training frontier.

In this setup, the *model parameters* are (q, p, c): properties of the AI system that are fixed at training time. The *decision-function parameters* are $(H, \bar{L}, c_V, c_S)$: properties of the agent's environment that determine how the prediction is used. The economic question is how the interaction between these two sets of parameters determines value.

2.1 Connection to Wald's framework

The structure above can be viewed as a reduced-form specialization of a Wald (1950)-type statistical decision problem. In Wald's setup, a decision maker chooses an action after observing the outcome of an experiment, defined as a signal about the state and characterized by a Markov kernel from the state space to the signal space. In our setting, the decision maker has access to multiple costly information-acquisition options: the AI query (cost c), verification (cost c_V, available only after a prediction), and solving (cost c_S, always available). Wald's basic insight still applies: the value of information is determined by the decisions it enables.

Sections 2–3.1, however, work with a reduced-form summary of the AI query rather than a full primitive signal structure. The model is summarized by (q, p, c), and later by the observable posterior π, instead of an explicit Markov kernel for model outputs. The Wald/Blackwell language should therefore be read as motivating the reduced form rather than as a full primitive characterization of the experiment. Literal Blackwell comparisons arise only once the posterior experiment is specified explicitly, as in Section 3.3.

2.2 Pairwise trust-verify comparison

Given the model triple (q, p, c), the two relevant payoffs from using the AI are:

$$\Pi_T = q\big[pH + (1-p)\bar{L}\big] + (1-q)B - c = B + a(H - \bar{L}) - q(B - \bar{L}) - c, \tag{2}$$

$$\Pi_V = q\big[pH + (1-p)B - c_V\big] + (1-q)B - c = B + a(H - B) - qc_V - c. \tag{3}$$

When trusting the AI, total payoff is the baseline payoff plus the additional payoff from correct recommendations, minus the loss from incorrect recommendations, minus the query cost; under verification, incorrect recommendations are replaced by the baseline at additional cost c_V. The difference is $\Pi_V - \Pi_T = q[(1-p)(B - \bar{L}) - c_V]$. This induces a natural *pairwise* partition between the two queried continuation actions: trust weakly dominates verify when $c_V \geq (1-p)(B - \bar{L})$, and verify strictly dominates trust when $c_V < (1-p)(B - \bar{L})$. This is only a comparison between trusting and verifying. The baseline action may still dominate both, so these regions do not by themselves determine adoption. For brevity, we will still call them the *trust* and *verify* regimes. As we will see in Section 2.6, this partition has a clean single-crossing property along the training frontier.

2.3 Canonical strategies

To make the reduced-form representation by (q, p, c) exact, we impose one further coarsening restriction.

Assumption 2 (Coarse error pooling). Conditional on the event that the model predicts, a failed verification reveals no payoff-relevant information beyond the fact that the recommendation is false. Equivalently, after observing a realized recommendation $\hat{\omega}$ and learning that $\hat{\omega} \neq \omega$, no feasible continuation action yields a higher expected payoff than reverting to the baseline, so the continuation value after failed verification is B.

Assumption 2 is what makes the queried AI a coarse experiment for the decision problem studied here: once the model predicts, the recommendation matters only through whether it is correct.[3]

Lemma 1 (Three canonical strategies)**.** *There exists an optimal policy that is one of:*

(1) ***Baseline:*** *Do not query; take the baseline value B.*

[3]This restriction is needed because the standing reduced-form assumption only says that, conditional on prediction, the realized recommendation carries no additional case-specific information about *whether* it is correct. That does not rule out the recommendation carrying information about *which states are likely when it is wrong*. If such residual information remains, then after a failed verification the agent may optimally take a contingent action that strictly improves on B.

(2) ***Trust:*** *Query; if the model predicts, act on the recommendation without verification; if it abstains, take the baseline. Payoff:* Π_T *in* (2).

(3) ***Verify:*** *Query; if the model predicts, verify, act only if verified correct, otherwise revert to baseline; if it abstains, take the baseline. Payoff:* Π_V *in* (3).

The model's economic value is $V(m) := \max\{B,\ \Pi_T,\ \Pi_V\}$.

Proof. If the model is not queried, the best achievable payoff is B by definition. Suppose the model is queried. On the abstention event, no new information is obtained, so the best continuation value is again B. On the prediction event, the agent either acts immediately on $\hat{\omega}$, yielding expected payoff $pH + (1-p)\bar{L}$, or pays c_V to verify. Under verification, if $\hat{\omega} = \omega$ the agent chooses $x = \omega$ and gets $H - c_V$; if $\hat{\omega} \neq \omega$, the agent learns only that the recommendation is false. By Assumption 2, if verification reveals that $\hat{\omega} \neq \omega$, the optimal continuation payoff is $B - c_V$. Thus any queried policy is weakly dominated by either trust or verify, and the overall optimum is $\max\{B, \Pi_T, \Pi_V\}$. □

This lemma allows us to simplify the agent decision problem and the model training problem: an agent will either query the model and trust it, query the model and try to verify its prediction before acting, or not use the model at all.

2.4 When to trust, when to verify, when to ignore

Proposition 1 (Autonomous use)**.** *Conditional on receiving a prediction, trusting without verification is optimal if and only if*

$$p \geq \tau := \frac{B - \bar{L}}{H - \bar{L}} \qquad \textit{and} \qquad c_V \geq (1-p)(B - \bar{L}).$$

Equivalently, trust must beat both the baseline and verification.

Proof. Conditional on a prediction, trust yields $pH + (1-p)\bar{L}$, verification yields $pH + (1-p)B - c_V$, and reverting immediately to the baseline yields B. Trust beats the baseline if and only if

$$pH + (1-p)\bar{L} \geq B \quad \iff \quad p \geq \frac{B - \bar{L}}{H - \bar{L}} = \tau.$$

Trust beats verification if and only if

$$pH + (1-p)\bar{L} \geq pH + (1-p)B - c_V \quad \iff \quad c_V \geq (1-p)(B - \bar{L}).$$

Since these are the only three continuation actions after a prediction, trust is optimal if and only if both inequalities hold. □

The second inequality is the pairwise trust-verify condition from Section 2.2. Trust, meaning autonomous AI use, is therefore used only when the model is accurate enough to clear a baseline threshold τ *and* verification is expensive relative to the expected loss avoided by checking.

Proposition 2 (Adoption)**.** *The model is used at all iff* $\max\{\Pi_T, \Pi_V\} \geq B$*. Equivalently:*

(i) When trust weakly dominates verify, adoption requires $c \leq a(H - \bar{L}) - q(B - \bar{L})$*.*

(ii) When verify strictly dominates trust, adoption requires $c \leq a(H - B) - qc_V$*.*

In particular, if $p < 1$, the model is not adopted whenever both

$$c_V > p(H - B) - \frac{c}{q} \qquad \textit{and} \qquad B - \bar{L} > \frac{p(H - B) - \frac{c}{q}}{1 - p}.$$

If $p = 1$, trust weakly dominates verify and adoption reduces to

$$c \leq q(H - B).$$

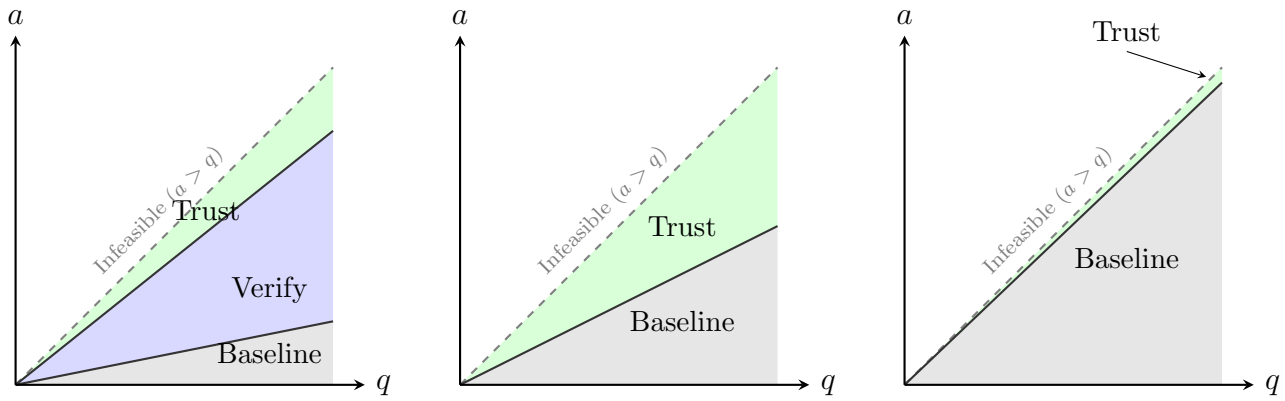
Proof. When trust weakly dominates verify, $\Pi_T \geq B$ rearranges to $c \leq a(H - \bar{L}) - q(B - \bar{L})$. When verify strictly dominates trust, $\Pi_V \geq B$ rearranges to $c \leq a(H - B) - qc_V$. Non-adoption is the conjunction. Continuity: $\Pi_T = \Pi_V$ at $c_V = (1 - p)(B - \bar{L})$. The final conditions follow by simple algebra. □

Consider Figure 1. In Panel A, where verification costs are low and downside risk from mistaken actions is minimal, the AI is widely used, and its predictions are verified for wide combinations of coverage and conditional correctness. In Panel B, where verification costs are higher, the verify region vanishes, and among adopted queried policies the AI is used only under trust. In Panel C, where downside risk is severe, even very accurate models are not used: in this case, 95% accuracy would not be sufficient because of the large downside risk and expense of verification.

2.5 Selective classification and Learning to Defer

There is a related problem in machine learning called *selective classification*: when should a model abstain rather than risk an incorrect prediction? The question arises whenever mistakes are more costly than declining to answer, precisely the condition $\bar{L} < B$ in our setting.[4]

[4]To date, most economic analyses of AI prediction in decision making have relied on models containing coverage *or* selective risk; see Gans (2025) for a discussion. Here these are combined as they capture two



Panel A: Low c_V **Panel B: High c_V** **Panel C: High $B - \bar{L}$**

Figure 1: Global optimal action regions in coverage-accuracy space (q, a). The dashed line $a = q$ bounds the feasible region since conditional correctness $p \leq 1$. **Panel A** ($H = 10, B = 5, \bar{L} = 0, c_V = 1, c = 0$) shows all three regimes emerging when verification costs are low ($c_V = 1$). **Panel B** ($H = 10, B = 5, \bar{L} = 0, c_V = 3, c = 0$) demonstrates that increasing verification cost ($c_V = 3$) causes the verify region to entirely collapse. **Panel C** ($H = 1, B = 0, \bar{L} = -20, c_V = 2, c = 0$) has severe downside cost ($\bar{L} = -20$) pushing the Trust threshold to $a/q > 20/21 \approx 0.952$. Consequently, the a model that always makes a prediction $q = 1$ and is accurate 95% of the time $a = p = .95$ falls into the Baseline region and is not adopted.

The classical answer, due to Chow (1970), is a posterior-threshold rule. In our notation: suppose the model can be summarized by a case-specific posterior probability of correctness s. Then trusting yields $sH + (1 - s)\bar{L}$, and the baseline yields B. Trust is optimal whenever

$$s \geq \tau := \frac{B - \bar{L}}{H - \bar{L}}. \tag{4}$$

This is immediate: the decision maker needs the expected payoff from trusting, $sH+(1-s)\bar{L}$, to exceed B, which rearranges to $s \geq (B - \bar{L})/(H - \bar{L})$.

This Chow threshold will reappear in the sequel as the slope of trust-regime iso-payoff curves and as the tangency condition for optimal training. This is not a coincidence, as the same tradeoff between errors and rejections drives both Chow's selective-classification problem and our decision problem, at least in the regime where all that can be done is trust the model or not use it. It does not accommodate the verify regime, however. In Chow's world, a misclassified instance simply incurs loss $\bar{L}$, with no option to pay c_V to check. Verification allows the agent to replace the expected incorrect-action payoff $(1 - p)\bar{L}$ with the expected baseline $(1 - p)B$, at a cost of c_V.

dimensions in real world machine-learning that interact at a design level.

Chow (1970) went further than the threshold rule via a simple characterization of the entire error-coverage frontier. Write $E(\tau)$ for the unconditional error rate - $q(1-p)$ in our notation - and $q(\tau)$ for the coverage under threshold τ. As τ rises (the model becomes more selective), coverage falls and errors fall. Chow proved that the error rate is determined by the coverage function alone via a Stieltjes integral:

$$E(\tau) = -\int_{\tau}^{1} (1-t)\, dq(t). \tag{5}$$

Each marginal unit of coverage lost (a small decrease in q) eliminates errors at rate $1-t$, the threshold at which those marginal instances were admitted. The most uncertain instances contribute the most errors per unit of coverage.

Two properties follow. First, the error-coverage frontier is convex: early selectivity yields large error reductions per unit of lost coverage, while further selectivity sacrifices increasingly confident outputs for diminishing error improvements. This convexity is the mathematical foundation for the diminishing returns to "playing it safe" that appear throughout our analysis. Second, the economic loss from using a free AI model under trust, relative to a perfect predictor, is

$$H - \Pi_T = (1 - q(\tau))(H - B) + E(\tau)(H - \bar{L}). \tag{6}$$

Thus the economic loss under trust reflects both abstention and classification error, unlike Chow's original formulation, which is organized directly around the error-coverage tradeoff.

Under verify, the frontier of feasible (q, p) pairs is unchanged since coverage and conditional correctness are properties of the AI, not of agent parameters. However, the economic loss becomes $H - \Pi_V = (H - B)(1 - qp) + qc_V$, which depends linearly on coverage through the verification cost qc_V. The Chow integral no longer suffices since it depends on how q feeds through to the verification cost qc_V rather than just the integrated error rate.

Note also that the convexity of the error-coverage frontier, established by (5), directly implies the concavity of the accuracy-coverage frontier: since unconditional accuracy is $a(q) = q - E(q)$ and E is convex, $a(\cdot)$ is concave. This provides the theoretical foundation for the concavity maintained in Assumption 3 below.

2.6 The training-possibilities frontier

Many modeling choices, including architecture, compute budget, fine-tuning strategy, decoding parameters, and confidence thresholds, can generate different accuracy-coverage pairs. In practice, a developer tunes these choices to move the model along a feasible set: a language model can be made to refuse uncertain queries (lowering q, raising p), a statistical model

can restrict inference to regions where identifying assumptions hold (the analog of selective prediction for traditional econometrics), or a classifier can threshold its softmax output at different levels.

Definition 1 (Training-possibilities frontier)**.** Given a model class and compute budget, the *training-possibilities frontier* is $A : [0,1] \to [0,1]$ where $A(q) := \sup\{a : (a,q) \text{ is achievable}\}$.

Note also that the convexity of Chow's error–coverage frontier motivates, but does not by itself prove, the strict concavity of the training frontier $A(\cdot)$ used below. We therefore impose strict concavity of A as a maintained regularity condition on the feasible accuracy–coverage set generated by a model class and training pipeline.

Assumption 3 (Frontier properties)*.* The frontier satisfies $A(0) = 0$, A is differentiable, and A is *strictly concave*.

Concavity reflects the natural ordering of instances by difficulty. Expanding coverage from zero first captures the easiest instances, yielding high marginal accuracy. Further expansion adds progressively harder instances with diminishing marginal accuracy. This shape is consistent with standard selective-prediction intuition and with empirical thresholding results in deep networks. Related theoretical work, such as Jones et al. (2021), studies conditions under which selective accuracy changes monotonically with the threshold under left-log-concavity of the margin distribution,[5] while El-Yaniv and Wiener (2010) provide the foundational risk–coverage framework and Geifman and El-Yaniv (2017) document this pattern empirically. We, therefore, treat Assumption 3 as an explicit maintained assumption rather than as a direct implication of any single cited theorem.

Since A is strictly concave with $A(0) = 0$, the ratio $p(q) = A(q)/q$ is strictly decreasing: expanding coverage lowers conditional correctness. Moreover, $A'(0) \leq 1$ (since $A(q) \leq q$ for all q, as $p \leq 1$), so the frontier lies weakly below the 45-degree line $a = q$ and has slope at most 1 everywhere. Figure 2 illustrates.

Lemma 2 (Single regime switch)**.** *The function $g(q) := (1-p(q))(B-\bar{L})$ is strictly increasing in q on $(0,1]$. For any $c_V > 0$:*

(i) If $g(q) \leq c_V$ for all $q \in (0,1]$, the entire frontier is in the trust regime.

(ii) If $g(q) \geq c_V$ for all $q \in (0,1]$, the entire frontier is in the verify regime.

(iii) Otherwise, there exists a unique $\tilde{q} \in (0,1)$ such that the frontier is in the trust regime for $q \leq \tilde{q}$ and in the verify regime for $q > \tilde{q}$.

[5]This is related to, but distinct from, strict concavity of the unconditional accuracy–coverage frontier $A(q)$ maintained here.

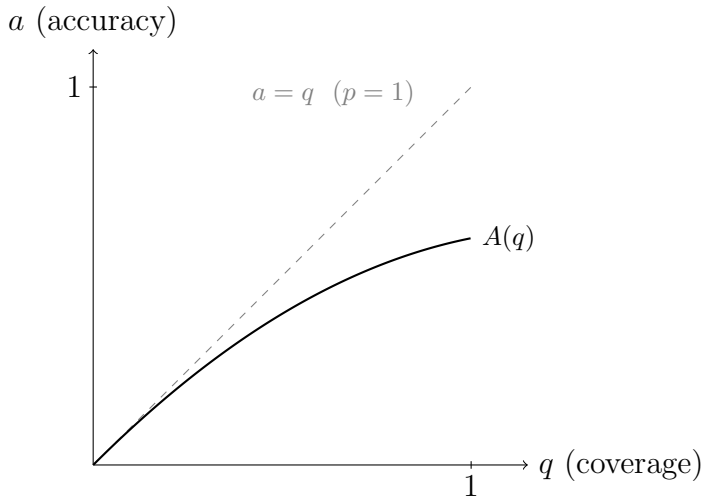


Figure 2: A concave training-possibilities frontier $A(q)$. The dashed line $a = q$ is the locus $p = 1$ (every prediction is correct). The frontier lies weakly below this line (since $p \leq 1$) and has slope at most 1 everywhere (by concavity through the origin). Expanding coverage from zero captures easy instances first, yielding high marginal accuracy; further expansion adds harder instances with diminishing returns.

Proof. Since $p(q) = A(q)/q$ is strictly decreasing (by strict concavity through the origin), $1 - p(q)$ is strictly increasing, and so is $g(q) = (1 - p(q))(B - \bar{L})$ since $B - \bar{L} > 0$. The trichotomy follows from continuity and strict monotonicity. □

As we move along the frontier toward higher coverage, the model becomes less conditionally accurate, so the incentive to verify grows monotonically. There is at most one switch from trust to verify, occurring where the expected benefit of verification exactly equals its cost.

3 Optimal Model Training

We have shown when agents will trust, verify, or not use a given model, and have introduced the idea of a model training frontier. We now ask how the developer should balance coverage and conditional accuracy when training their model to be used by humans. Because users in this section observe only a pooled prediction about the true state, the developer's choice of coverage on $A(q)$ effectively determines how selective the model will be in use.

To work up to the developer optimum, consider first the choice of coverage q and conditional correctness p, *holding fixed* overall accuracy a. Along any given isoaccuracy curve, in either the trust or verify regimes, payoffs are strictly increasing in p. That is, if it were possible to train a model with the same accuracy but higher conditional correctness and hence lower coverage, all agents would strictly prefer this.

Proposition 3 (Iso-accuracy monotonicity)**.** *Fix overall accuracy $a \in (0, 1]$ and consider models with $q > 0$ and $p = a/q$.*

(i) Under trust, Π_T is strictly increasing in p along $a = qp$.

(ii) Under verify, if $c_V > 0$ then Π_V is strictly increasing in p along $a = qp$.

Proof. Fix a and substitute $q = a/p$ into (2): $\Pi_T(p) = B + a(H - \bar{L}) + (a/p)(\bar{L} - B) - c$. Then $d\Pi_T/dp = (a/p^2)(B - \bar{L}) > 0$ since $B > \bar{L}$ by Assumption 1. For verify: $\Pi_V(p) = B + a(H - B) - (a/p)c_V - c$, so $d\Pi_V/dp = (a/p^2)c_V > 0$. □

Corollary 1 (When selectivity matters most)**.** *The marginal value of increasing p along the iso-accuracy curve is $(a/p^2)(B - \bar{L})$ under trust and $(a/p^2)c_V$ under verify. The gain from selectivity grows linearly in $(B - \bar{L})$ and in c_V. That is, the worse the consequences of acting on an incorrect prediction, either because verification is costly or consequences of wrong actions are severe, the more valuable it is for the AI to know when it does not know.*

3.1 The optimal model on the frontier

We now ask what the optimal model training target is, given an accuracy-coverage frontier $A(q)$. In the coarse-output environment, this is the core design problem: if users cannot distinguish predictions once the model speaks, then the developer must choose how much selectivity to build into the system ex ante. Since $A(q)$ is strictly concave, expanding coverage lowers conditional correctness. Along the upward-sloping part of the frontier, higher unconditional accuracy is obtained only by accepting lower conditional correctness. We know from the previous proposition that payoffs are strictly increasing in p holding a fixed. The optimal point on the coverage-accuracy frontier therefore is where the marginal gain from increasing accuracy is exactly equal to the marginal cost of additional errors. This will be a tangency condition on the accuracy-coverage frontier for each of the decision problem regimes.

Proposition 4 (Optimal model training target)**.** *Fix decision-function parameters and a differentiable, strictly concave frontier $A(\cdot)$. Let*

$$\tau := \frac{B - \bar{L}}{H - \bar{L}}, \qquad \sigma := \frac{c_V}{H - B}.$$

Let I_T and I_V denote the feasible trust and verify intervals from Lemma 2; if one regime is absent, the corresponding interval is empty.

(i) The regime-specific optimization problems are concave:

$$\max_{q \in I_T} \Big\{ B + A(q)(H - \bar{L}) - q(B - \bar{L}) - c \Big\},$$

$$\max_{q \in I_V} \Big\{ B + A(q)(H - B) - q c_V - c \Big\}.$$

If the regime-specific maximizer is interior in its interval, then

$$A'(q_T^\star) = \tau, \qquad A'(q_V^\star) = \sigma.$$

If the tangency point lies outside the relevant regime interval, the regime-specific maximizer is the appropriate boundary point of that interval.

(ii) Let $q^a \in \arg\max_q A(q)$. If q^a is an interior maximizer of A, then q^a can be economically optimal only in the verify regime with $c_V = 0$. If q^a is a boundary maximizer (in particular $q^a = 1$ with $A'(1) > 0$), then q^a is optimal only if the frontier is at least as steep as the governing iso-payoff slope at q^a: that is, $A'(q^a) \geq \tau$ if q^a lies in the trust regime, or $A'(q^a) \geq \sigma$ if q^a lies in the verify regime.

(iii) On any regime where the optimum is interior, holding H and B fixed, $q_T^\star$ decreases as $\bar{L}$ falls and $q_V^\star$ decreases with c_V.

(iv) The global optimum is obtained by comparing the regime-specific candidates together with the relevant boundary points of I_T and I_V; when both regimes are present, this includes the switch point $\tilde{q}$ from Lemma 2. Overall adoption then compares the best such models with the baseline value B.

Proof. See Appendix 6.1. □

When users cannot distinguish good predictions from bad ones once the model speaks, the developer should train the model to solve the selection problem on their behalf. In high-downside or high-verification-cost environments, this means training for selective competence rather than maximal coverage; when verification is cheap and effective, it means training for broader coverage and making more cases available for review; when verification is expensive but the harms of mistakes are low, it again means training for broader coverage and hence unconditional accuracy. Operationally, the relevant design levers are abstention behavior, refusal thresholds, and other choices that move the deployed system along $A(q)$.

There are a few interpretations of this result, geometric and in terms of traditional machine learning results. First, geometry. Hold fixed some payoff and consider iso-payoff curves in (a, q) space:

$$\left.\frac{da}{dq}\right|_{\Pi_T} = \frac{B - \bar{L}}{H - \bar{L}} = \tau, \qquad \left.\frac{da}{dq}\right|_{\Pi_V} = \frac{c_V}{H - B}. \tag{7}$$

The trust-regime iso-payoff slope is exactly the Chow threshold τ and the verify-regime iso-payoff slope is exactly the verification cost per unit of potential gain σ. That is, we consider the relationship between additional accuracy and additional coverage that keeps payoffs constant for a given regime. We then compare this to the slope of the accuracy-coverage frontier $A'(q)$, which is the marginal gain from additional coverage. The optimal training target is then the point where these two slopes are equal.

The tangency condition has a clean interpretation that connects back to Chow. In the trust regime, $A'(q^\star) = \tau$ says we should expand coverage until the marginal predictions are exactly at the posterior threshold where the decision maker is indifferent between trusting and taking the baseline payoff. Instances below this threshold are not worth predicting on as they would be rejected under Chow's rule. The training condition *is* the Chow condition applied at the margin. In the verify regime, $A'(q^\star) = c_V/(H - B)$ says we ought expand coverage until the marginal accuracy gain from one more predicted instance equals the verification cost per unit of potential gain. It should be clear from the mechanics of this proposition that one can additional regimes of human behavior, such as data augmentation, which will introduce additional regime regions and tangency conditions: our results do not depend on some unusual modeling of verification, though we see that as a key aspect of human-AI interaction.

Figure 3 illustrates these tangency conditions.

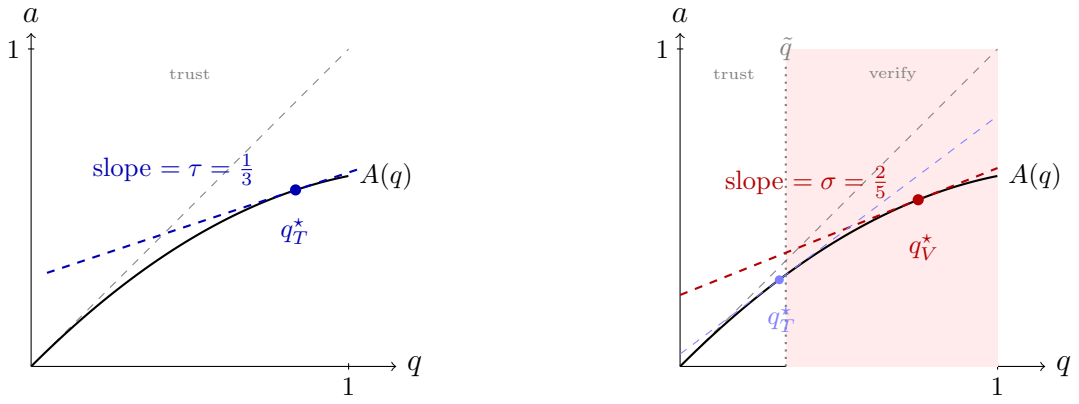


Figure 3: Optimal model on the frontier $A(q) = q - 0.4q^2$. **(a)** A low-stakes environment ($H = 10$, $\bar{L} = -5$, $c_V = 4$, $B = 0$, $c = 0$): verification is expensive enough that the entire frontier lies in the trust regime. The tangency at $\tau = \frac{1}{3}$ yields $q_T^\star \approx 0.83$, favoring broad coverage. **(b)** A high-stakes environment ($H = 10$, $\bar{L} = -30$, $c_V = 4$, $c = 0$, $B = 0$): the regime boundary $\tilde{q} \approx 0.33$ divides the frontier. The trust tangency (faint blue, $\tau = \frac{3}{4}$) occurs at low coverage ($q_T^\star \approx 0.31$); the verify tangency (red, $\sigma = \frac{2}{5}$) occurs at $q_V^\star = 0.75$. The verify tangency dominates, delivering higher value at moderate coverage with costly but worthwhile verification.

Note that even though the profit function is continuous across regimes, the optimal coverage $q^\star$ can jump discontinuously as the economic environment varies. Because the global optimum is $\max\{\Pi_T(q_T^\star),\ \Pi_V(q_V^\star),\ B\}$, and the regime-specific candidates $q_T^\star$ and $q_V^\star$ generically differ, a small change in (τ, σ) that tips the global optimum from one regime to the other produces a discrete jump in optimal model training.

To illustrate, consider a medical diagnostic AI where the baseline value of abstaining is $B = 0$, the payoff for a correct intervention is $H = 100$, and the expected loss of an incorrect intervention is severe, $\bar{L} = -500$. Assume the query is free ($c = 0$) and the training-possibilities frontier takes the quadratic form $A(q) = q - 0.4q^2$. In the trust regime, the optimal training target depends only on the downside risk, with slope

$$\tau = \frac{B - \bar{L}}{H - \bar{L}} = \frac{500}{600} \approx 0.833.$$

The tangency condition $A'(q_T^\star) = \tau$ yields a highly selective optimal coverage of $q_T^\star \approx 0.21$, generating a payoff of $\Pi_T \approx 10.4$.

However, if verification is relatively cheap, the agent will prefer the verify regime. At $c_V = 40$, the verify slope is

$$\sigma = \frac{c_V}{H - B} = \frac{40}{100} = 0.40.$$

The tangency condition $A'(q_V^\star) = \sigma$ yields a much more aggressive optimal coverage of $q_V^\star = 0.75$, generating a payoff of $\Pi_V = 22.5$. Because $\Pi_V > \Pi_T$, the globally optimal model should be trained to target $q^\star = 0.75$.

As the cost of verification rises, the verify regime becomes less attractive, and the model must become more selective to compensate. At $c_V = 58$, optimal coverage in the verify regime smoothly falls to $q_V^\star \approx 0.53$, and the payoff drops to $\Pi_V \approx 11.0$. But as c_V crosses a critical threshold near $c_V = 59$, the payoff from verifying falls below the trust payoff of approximately 10.4. At this exact point, the agent abandons verification entirely. The globally optimal training target discontinuously plunges from $q^\star \approx 0.51$ down to $q^\star \approx 0.21$.

3.2 Heterogeneity and optimal models

We now relax the representative-agent assumption because one of the central economic facts about AI is that a single costly model is typically deployed to many users. This extension isolates the cost of one-size-fits-all model design. In the reduced-form setting, users cannot re-sort predictions for themselves, so disagreement about downside risk or verification cost becomes disagreement about what should be trained in the first place.

We showed (Proposition 4) that the optimal model depends on the decision-function pa-

rameters through τ and $c_V/(H-B)$. Because the optimal training target depends strictly on agent-specific parameters, a developer forcing a single model onto a heterogeneous population creates inevitable welfare losses. A direct consequence is that two models can reverse in ranking across environments, and the model (q, p, c) does not induce an environment-invariant ranking: optimal model selection requires knowledge of the agent's decision-function parameters.

Remark 1 (No environment-invariant ranking). Two models $m_1 = (q_1, p_1, c)$ and $m_2 = (q_2, p_2, c)$ can satisfy $V(m_2; e_1) > V(m_1; e_1)$ but $V(m_1; e_2) > V(m_2; e_2)$ for two environments e_1, e_2. A model with high coverage and moderate conditional correctness dominates when errors are mild; a model with low coverage and high conditional correctness dominates when errors are very costly. This contrasts with the Blackwell (1951, 1953) ordering, which ranks experiments universally, and reflects the fact that a reduced-form pair (q, p) does not determine a complete information structure.

Since agents will disagree about which model is best, let us consider which types of agents are most hurt by population-level model selection. Consider a world where a single model is trained. The planner will choose the coverage-accuracy tradeoff that maximizes population welfare. We can then ask which types of agents are most hurt by this choice.

Proposition 5 (Population-optimal training with mixed trust and verify agents)**.** *Let μ be a population distribution over agent types θ with common H, and suppose the developer ships a single model $m = (q, p, c)$ with q, p chosen on $A(\cdot)$. For type θ, define*

$$\begin{aligned} W_\theta^0(q) &:= B_\theta, \\ W_\theta^T(q) &:= B_\theta + A(q)(H - \bar{L}_\theta) - q(B_\theta - \bar{L}_\theta) - c, \\ W_\theta^V(q) &:= B_\theta + A(q)(H - B_\theta) - q c_{V,\theta} - c, \end{aligned}$$

and let

$$W_\theta(q) := \max\{W_\theta^0(q),\, W_\theta^T(q),\, W_\theta^V(q)\}.$$

Population welfare is

$$\mathcal{W}(q) := \int W_\theta(q)\, d\mu(\theta).$$

Fix an open interval $I \subseteq (0, 1)$ on which, for μ-a.e. θ, the maximizing branch in $W_\theta(q)$ is unique and constant in q. Let T_I denote the set of types whose maximizing branch on I is trust, and V_I the set of types whose maximizing branch on I is verify.

(i) On I,

$$\mathcal{W}(q) = C_I + A(q)\left[\int_{T_I}(H-\bar{L}_\theta)\,d\mu(\theta)+\int_{V_I}(H-B_\theta)\,d\mu(\theta)\right]-q\left[\int_{T_I}(B_\theta-\bar{L}_\theta)\,d\mu(\theta)+\int_{V_I}c_{V,\theta}\,d\mu(\theta)\right]$$

for a constant C_I. Hence $\mathcal{W}$ is concave on I, and strictly concave whenever $\mu(T_I \cup V_I) > 0$.

(ii) If $\mu(T_I \cup V_I) > 0$, any interior maximizer $q_I^\dagger \in I$ satisfies

$$A'(q_I^\dagger) = \frac{\int_{T_I}(B_\theta - \bar{L}_\theta)\,d\mu(\theta) + \int_{V_I} c_{V,\theta}\,d\mu(\theta)}{\int_{T_I}(H - \bar{L}_\theta)\,d\mu(\theta) + \int_{V_I}(H - B_\theta)\,d\mu(\theta)}. \tag{8}$$

(iii) If μ has finite support, the global optimum is obtained by comparing the finitely many interval-specific candidates from (8) with the finitely many boundary points at which some type switches between baseline, trust, and verify.

In the special cases $V_I = \varnothing$ or $T_I = \varnothing$, (8) reduces to the all-trust and all-verify formulas.

Corollary 2 (Quadratic regret on a fixed branch)**.** *Fix type θ, and let*

$$\Delta_\theta := W_\theta(q_\theta^\star) - W_\theta(q^\dagger) \geq 0.$$

Suppose the maximizing branch for type θ is the same at $q_\theta^\star$ and $q^\dagger$, and on the interval between them, and suppose $q_\theta^\star$ is interior on that branch. If A is twice differentiable on this interval, then

$$\Delta_\theta = -\frac{1}{2}\,W''_{\theta,b}(\tilde{q}_\theta)\,(q^\dagger - q_\theta^\star)^2 \tag{9}$$

for some $\tilde{q}_\theta$ between $q^\dagger$ and $q_\theta^\star$, where $b \in \{T, V\}$ is the common branch. If the maximizing branch changes between $q_\theta^\star$ and $q^\dagger$, one still has $\Delta_\theta \geq 0$, but the exact quadratic representation need not hold.

Proof. See Appendix 6.2. □

Regret is quadratic in the distance between the agent's ideal coverage and the deployed coverage. Agents with unusually high $B_\theta - \bar{L}_\theta$ are especially hurt in the trust region: they want lower coverage and higher conditional correctness than the population optimum provides. Agents with unusually high $c_{V,\theta}$ are especially hurt in the verify region for the same reason. Symmetrically, unusually low-risk agents are hurt when the common model is too selective. If the population is dominated by high-downside, high-verification-cost agents, the

deployed point has low coverage, and agents who don't mind the risk or who could cheaply verify bad predictions see lower welfare.

3.3 State-specific signals

We now relax the coarse-information assumption by allowing the model to emit scores, confidence reports, or other side information that helps the user assess reliability case by case. This changes the relevant design margin. In Sections 2-3.2, the developer solves the selection problem largely at training time by choosing coverage on $A(q)$. Once the model reveals calibrated state-specific reliability, part of that selection is decentralized: the developer chooses what reliability information to reveal, and users screen predictions for themselves. The question becomes when value is created by selective training and when it is created by informative, calibrated reporting.[6]

Formally, let the model issue a recommendation $\hat{\omega}$ together with an observable signal from which the agent can compute a posterior

$$\pi := \Pr(\hat{\omega} = \omega \mid \text{signal}), \qquad \pi \in [0, 1].$$

Let μ denote the distribution of π across queried cases, with continuous density $f > 0$ on a support $[\underline{\pi}, \bar{\pi}] \subseteq [0, 1]$. For any measurable $E \subseteq [0, 1]$, define coverage $q(E) := \mu(E)$ and unconditional accuracy mass $a(E) := \int_E \pi \, d\mu(\pi)$, so that average conditional correctness on E is $p(E) = a(E)/q(E)$.

This framework nests two polar cases. If the posterior is not observed, so that all acted-on cases share a pooled posterior $\pi \equiv p$, we recover the reduced-form model of Sections 2-3.1. If the posterior is observed as a calibrated score s with $s = \Pr(\hat{\omega} = \omega \mid s)$, the agent conditions directly on instance-level reliability.

The pointwise continuation payoffs after observing posterior π are:

$$\begin{aligned} U_T(\pi) &:= \pi H + (1 - \pi)\bar{L}, \\ U_V(\pi) &:= \pi H + (1 - \pi)B - c_V, \\ U_B &:= B. \end{aligned}$$

All three are affine in π, with U_T having slope $H - \bar{L}$, U_V having slope $H - B$, and U_B being

[6] To our knowledge, Autor et al. (2025) are the first to study post-processing calibration for preferences rather than changes in model training that incorporate population-level preference information. See also Gans (2026b) for a general treatment. Gans (2026a) examines how calibration can improve settings in which model accuracy is uneven (or jagged) but that information is not available to users.

flat. Pairwise comparisons are governed by the same objects as in Section 2.2:

$$U_T(\pi) \geq U_B \quad \iff \quad \pi \geq \tau := \frac{B - \bar{L}}{H - \bar{L}}, \tag{10}$$

$$U_V(\pi) \geq U_B \quad \iff \quad \pi \geq \sigma := \frac{c_V}{H - B}, \tag{11}$$

$$U_T(\pi) \geq U_V(\pi) \quad \iff \quad c_V \geq (1 - \pi)(B - \bar{L}). \tag{12}$$

Proposition 6 (Autonomous use with posteriors)**.** *Conditional on observing posterior π, trusting without verification is optimal if and only if*

$$\pi \geq \tau := \frac{B - \bar{L}}{H - \bar{L}} \qquad \textit{and} \qquad c_V \geq (1 - \pi)(B - \bar{L}).$$

Equivalently, trust must beat both the baseline and verification at that posterior.

Proof. Trust is optimal at posterior π if and only if $U_T(\pi) \geq U_B$ and $U_T(\pi) \geq U_V(\pi)$. By (10), the first condition is equivalent to $\pi \geq \tau$. By (12), the second is equivalent to $c_V \geq (1 - \pi)(B - \bar{L})$. □

This specializes Proposition 1. In the reduced-form model, $\pi \equiv p$, so trust requires both $p \geq \tau$ and $c_V \geq (1 - p)(B - \bar{L})$. With calibrated scores, high-score predictions may be trusted, intermediate-score predictions may be verified, and low-score predictions may be ignored.

Proposition 7 (Adoption with posteriors)**.** *Define the pointwise value added by a prediction at posterior π:*

$$\phi(\pi) := \max\big\{0,\ \pi(H - \bar{L}) - (B - \bar{L}),\ \pi(H - B) - c_V\big\}. \tag{13}$$

Under the optimal pointwise continuation action, the model is adopted if and only if $\mathbb{E}[\phi(\pi)] \geq c$.

The optimal action partitions the posterior space:

(a) If $c_V \geq (B - \bar{L})(H - B)/(H - \bar{L})$, verification is never optimal. The rule is Chow's: trust if $\pi \geq \tau$, baseline otherwise.

(b) If $c_V < (B - \bar{L})(H - B)/(H - \bar{L})$, define

$$\kappa := 1 - \frac{c_V}{B - \bar{L}}.$$

Then $\sigma < \tau < \kappa$, and the optimal rule is three-way:

$$\textit{baseline if } \pi < \sigma, \qquad \textit{verify if } \sigma \le \pi < \kappa, \qquad \textit{trust if } \pi \ge \kappa.$$

Proof. The expected payoff from querying is $W = B + \mathbb{E}[\mathbf{1}\{\text{predict}\} \cdot \phi(\pi)] - c$, where the agent takes the payoff-maximizing action at each π. Adoption requires $W \ge B$, giving the displayed condition.

For the partition, compare the three affine functions U_B, U_T, and U_V. Verification is never optimal if and only if the trust–verify crossing lies weakly below the trust–baseline crossing, that is,

$$1 - \frac{c_V}{B - \bar{L}} \le \frac{B - \bar{L}}{H - \bar{L}},$$

which rearranges to $c_V \ge (B - \bar{L})(H - B)/(H - \bar{L})$. In that case the upper envelope is simply $\max\{U_B, U_T\}$, so the Chow rule applies.

If $c_V < (B - \bar{L})(H - B)/(H - \bar{L})$, let $\kappa := 1 - c_V/(B - \bar{L})$. Then $\sigma < \tau < \kappa$. For $\pi < \sigma$, verification is below baseline, and trust is also below baseline because $\pi < \tau$; hence baseline is optimal. For $\sigma \le \pi < \kappa$, verification beats baseline by definition of σ and beats trust by definition of κ. For $\pi \ge \kappa$, trust beats verification, and since $\kappa > \tau$, it also beats the baseline. Thus the stated three-way partition is optimal. □

Proposition 8 (Iso-accuracy monotonicity with posteriors)**.** *Fix unconditional accuracy mass $a \in (0, 1]$. Among all measurable prediction sets $E \subseteq [0, 1]$ with $a(E) = \int_E \pi \, d\mu(\pi) = a$:*

(i) Under trust,

$$W_T(E) = B + a(H - \bar{L}) - q(E)(B - \bar{L}) - c,$$

which is strictly increasing in average conditional correctness $p(E) = a/q(E)$.

(ii) Under verify, if $c_V > 0$,

$$W_V(E) = B + a(H - B) - q(E)\, c_V - c,$$

which is strictly increasing in $p(E)$.

(iii) The payoff-maximizing prediction set at fixed a is the upper tail of π: predict on the highest-posterior cases first.

Proof. Parts (i)-(ii): substitute $a(E) = a$ and note that both payoffs are strictly decreasing in coverage $q(E)$ holding a fixed ($B - \bar{L} > 0$ and $c_V > 0$), hence strictly increasing in

$p(E) = a/q(E)$. Part (iii): among measurable sets with fixed $\int_E \pi \, d\mu = a$, the set that minimizes $\mu(E)$ is the upper tail of π. This is the one-dimensional rearrangement inequality: to achieve a given integral of a nonnegative function, one uses as little measure as possible by selecting the largest values first.[7] □

This is the within-model analogue of Proposition 3. At fixed unconditional accuracy mass, value rises with conditional correctness, so optimal deployment selects the highest-posterior cases first. Thresholding on scores is therefore the micro-foundation for the reduced-form coverage-accuracy tradeoff.

To formulate the training analogue of Proposition 4, define the *induced posterior frontier*

$$\mathcal{A}(q) := \sup \left\{ \int_E \pi \, d\mu(\pi) : \; E \subseteq [0,1] \text{ measurable with } \mu(E) = q \right\}. \tag{14}$$

If μ is atomless, then $\mathcal{A}$ is concave, differentiable where the upper quantile function $Q(q)$ of π is continuous, and $\mathcal{A}'(q) = Q(q)$: the slope of the posterior frontier at coverage q equals the posterior correctness of the marginal admitted case.

Proposition 9 (Optimal training target with posteriors)**.** *Fix decision-function parameters and suppose $\mathcal{A}(\cdot)$ is differentiable and strictly concave.*

(i) The trust-only and verify-only problems

$$\max_{q \in [0,1]} \Big\{ B + \mathcal{A}(q)(H - \bar{L}) - q(B - \bar{L}) - c \Big\}, \quad \max_{q \in [0,1]} \Big\{ B + \mathcal{A}(q)(H - B) - q\, c_V - c \Big\}$$

are concave. Interior optima satisfy

$$\mathcal{A}'(q_T^\star) = \tau = \frac{B - \bar{L}}{H - \bar{L}}, \qquad \mathcal{A}'(q_V^\star) = \sigma = \frac{c_V}{H - B}.$$

Since $\mathcal{A}'(q) = Q(q)$ is the posterior of the marginal admitted case, these conditions say: expand coverage until the marginal prediction has the posterior at which the agent is indifferent between acting and taking the baseline.

(ii) If the agent observes π and applies the pointwise-optimal action from Proposition 7, the developer's value under a threshold deployment rule $\{\pi \geq t\}$ is

$$V(t) = B + \int_t^{\bar{\pi}} \phi(\pi) f(\pi) \, d\pi - c.$$

[7] Upper-tail selection is unique when μ is atomless. If μ has an atom at the marginal posterior, multiple optimal sets exist that differ on the atom, but the optimized payoff is the same.

Since $V'(t) = -\phi(t)f(t)$, and $\phi(\pi) = 0$ for $\pi \leq s^\star := \min\{\sigma, \tau\}$ while $\phi(\pi) > 0$ for $\pi > s^\star$, every feasible threshold $t \leq s^\star$ attains the same maximal value. The most selective optimal threshold is therefore

$$t^\star = s^\star = \min\{\sigma, \tau\}.$$

Equivalently, the model should predict on all cases with $\pi \geq t^\star$; lowering the threshold further is payoff-irrelevant because the agent ignores those additional predictions.

Proof. See Appendix 6.3. □

Part (i) reproduces Proposition 4: when posteriors are not observed, the developer still internalizes the coverage-accuracy tradeoff through a tangency condition. Part (ii) is the new case. With observable posteriors, low-π predictions are screened out by the agent, so the developer need only include all cases above $s^\star = \min\{\sigma, \tau\}$. Selection is therefore centralized at training time in the reduced-form model and decentralized at use when posteriors are observed. If the model reports a calibrated score $s = \pi$, the implementation is immediate: deploy on all scores above $\min\{\sigma, \tau\}$ and let the agent apply the three-way partition of Proposition 7 score by score. This result allows us to build to a Blackwell-ordering.

Proposition 10 (Heterogeneity with observable posteriors)**.** *Suppose agents observe the posterior π and the developer chooses a common threshold t on the posterior frontier $\mathcal{A}(\cdot)$. Let $s^\star_\theta := \min\{\sigma_\theta, \tau_\theta\}$ denote type θ's lowest posterior at which some action beats the baseline. Then:*

(i) Define

$$t^\dagger := \inf_\theta s^\star_\theta.$$

Every feasible threshold $t \leq t^\dagger$ is population-optimal. In particular, $t^\dagger$ is the most selective population-optimal threshold.

(ii) Every type achieves their first-best value within the given posterior experiment at any population-optimal threshold: if $t \leq t^\dagger$, then

$$V_\theta(t) = V^\star_\theta \qquad \text{for all } \theta.$$

In particular,

$$\Delta_\theta := V^\star_\theta - V_\theta(t^\dagger) = 0 \qquad \text{for all } \theta.$$

Proof. At threshold $t^\dagger$, type θ faces predictions with $\pi \in [t^\dagger, \bar{\pi}]$ and applies their own pointwise-optimal rule. For $\pi < s^\star_\theta$, type θ takes the baseline ($\phi_\theta(\pi) = 0$), so including

those predictions is costless. For $\pi \geq s^\star_\theta$, type θ acts optimally. Hence every type achieves exactly the payoff they would obtain under their own first-best threshold $s^\star_\theta$ within this posterior experiment. Lowering t below $t^\dagger$ adds predictions that all types ignore, so aggregate value is unchanged; raising t above $t^\dagger$ removes predictions that at least one type values, so no such threshold can be optimal. □

This contrasts sharply with Corollary 2. In the reduced-form model, a common coverage level imposes welfare losses on types whose ideal coverage differs from the deployed one. With observable posteriors, users recover their preferred effective coverage by ignoring low-posterior predictions. The quadratic regret in the coarse model is thus the welfare cost of coarse information.

Corollary 3 (Blackwell ordering with observable posteriors)**.** *Suppose two models both produce observable posteriors and model A's posterior experiment Blackwell-dominates model B's. Then $V_\theta(A) \geq V_\theta(B)$ for every agent type θ.*

Proof. If B's posterior is a garbling of A's, any posterior-contingent deployment rule feasible under B can be replicated under A by first applying the garbling and then the same rule. Since ϕ_θ in (13) is convex in π (as the pointwise maximum of affine functions), Jensen's inequality gives $\mathbb{E}_A[\phi_\theta(\pi)] \geq \mathbb{E}_B[\phi_\theta(\pi)]$ for every type θ. □

In the coarse environment of our baseline case, where the model reports a prediction, value in human-AI interaction comes from selectivity in the model training. With observable calibrated posteriors, value comes from making uncertainty visible and actionable, so that different users can impose different trust and verification thresholds on the same model. Training therefore shifts away from a scalar accuracy target and toward informative, calibrated posterior signals that the interface exposes cheaply enough for users to act on.

The design rule is therefore to invest in richer posterior reporting when users can cheaply observe calibrated scores and act on them; otherwise invest in selective training. Why might "coarse predictions" therefore be used? First, observing π may itself be costly. If inspection costs $c_E > 0$ per prediction, posterior-contingent action is worthwhile only when its gain exceeds c_E. Users for whom lookup is too costly face the pooled posterior p and revert to the reduced-form setting. Second, reported scores may be noisy or miscalibrated. Then the pointwise partition need not be optimal, and the Blackwell ordering, zero-regret result, and unanimous ranking can fail. Miscalibration is pervasive in current systems (Guo et al., 2017). Third, even perfect posteriors decentralize threshold choice within a model, not model choice across frontiers. Two posterior frontiers $\mathcal{A}_1(\cdot)$ and $\mathcal{A}_2(\cdot)$ induce different posterior

distributions, and agents with different iso-payoff slopes may disagree about which frontier is better. These frictions are common in practice, especially for large language models.

3.4 Monopoly training

We now move from planner design to market provision. This extension matters because AI models are costly to train and are therefore often supplied as common products by a small number of firms rather than customized user by user. The question is whose selectivity demand gets built into the common model once a firm chooses both the training target and the price. The previous subsection showed that heterogeneous agents disagree about the optimal model because they disagree about the relevant iso-payoff slope. For the monopoly problem, that slope is the natural one-dimensional summary of type.

To see this, write the gross value of a model at coverage q using the frontier $A(q) = qp(q)$. Under trust,

$$W_\theta^T(q) - B + c = A(q)(H - \bar{L}_\theta) - q(B - \bar{L}_\theta) = (H - \bar{L}_\theta)\big[A(q) - \tau_\theta q\big], \tag{15}$$

where

$$\tau_\theta := \frac{B - \bar{L}_\theta}{H - \bar{L}_\theta}.$$

Under verify,

$$W_\theta^V(q) - B + c = A(q)(H - B) - q c_{V,\theta} = (H - B)\big[A(q) - \sigma_\theta q\big], \tag{16}$$

where

$$\sigma_\theta := \frac{c_{V,\theta}}{H - B}.$$

Conditional on a regime, the multiplicative term in (15) or (16) is positive and affects the *level* of willingness to pay but not the buyer's preferred coverage. The desired model solves

$$A'(q) = \tau_\theta \qquad \text{in trust,} \qquad A'(q) = \sigma_\theta \qquad \text{in verify.}$$

Hence different combinations of downside risk and verification cost are equivalent for training whenever they generate the same relevant slope. The joint distribution of $(\bar{L}_\theta, c_{V,\theta})$ therefore induces a distribution of *effective slopes*, which is the object that matters for the choice of q.

To keep the monopoly section focused on the distortion in the training target rather than on separate level differences in willingness to pay, we now take that induced slope distribution

as primitive. Thus a buyer is indexed by a scalar

$$s \in [\underline{s}, \bar{s}] \subset (0, A'(0)),$$

distributed according to a continuously differentiable cdf G with density $g > 0$, and obtains gross surplus

$$u(s, q) = A(q) - sq \tag{17}$$

from a model trained at coverage q. Low-s buyers tolerate high coverage and lower conditional correctness; high-s buyers prefer more selective models. This reduction is exact for the buyer's preferred model conditional on a regime, and deliberately suppresses separate heterogeneity in value levels so that the monopoly section isolates the coverage distortion as cleanly as possible.

Proposition 11 (Monopoly training optimum)**.** *Let buyer types be distributed according to G on $[\underline{s}, \bar{s}]$, with density $g > 0$, and suppose the relevant optima deliver nonnegative gross surplus for the types served.*

(i) If the model is made available to all buyers at zero price, welfare is

$$\mathcal{W}(q) = \int_{\underline{s}}^{\bar{s}} \big[A(q) - sq\big]\, dG(s) = A(q) - q\,\mathbb{E}[s].$$

Hence $\mathcal{W}$ is strictly concave. If $A'(1) < \mathbb{E}[s]$, the unique socially optimal coverage is interior and satisfies

$$A'(q^S) = \mathbb{E}[s]. \tag{18}$$

If $A'(1) \geq \mathbb{E}[s]$, then $\mathcal{W}$ is increasing on $[0, 1]$ and $q^S = 1$.

(ii) A monopolist chooses a single coverage level q and a uniform price r. Since $u(s, q)$ is strictly decreasing in s, there exists a cutoff type $\hat{s}$ such that all buyers with $s \leq \hat{s}$ purchase and all buyers with $s > \hat{s}$ do not. Writing the price as

$$r = A(q) - \hat{s}q,$$

profit can be expressed as

$$\Pi(q, \hat{s}) = \big[A(q) - \hat{s}q\big]\, G(\hat{s}). \tag{19}$$

(iii) Any interior monopoly optimum $(q^M, \hat{s}^M)$ satisfies

$$A'(q^M) = \hat{s}^M \tag{20}$$

and

$$A(q^M) - \hat{s}^M q^M = q^M \frac{G(\hat{s}^M)}{g(\hat{s}^M)}. \tag{21}$$

(iv) When both q^M and q^S are interior, A' is strictly decreasing, so

$$q^M > q^S \iff \hat{s}^M < \mathbb{E}[s]. \tag{22}$$

The proposition gives a simple interpretation of the monopoly distortion. The planner targets the *population-average* selectivity demand, $\mathbb{E}[s]$. Equation (20) says that the monopolist trains for the *marginal buyer's* demand for selectivity. When the firm sells mainly to routine, low-s users and excludes high-s users, the deployed model has too much coverage and too little conditional correctness relative to the social optimum. When instead the firm serves the whole market, the marginal buyer has high loss and high verification costs, and the model is too conservative. Monopoly distorts training through the same object as the planner's problem, but around the wrong user. A platform selling one common model will train too aggressively when its paying base is skewed toward routine, low-risk users and too conservatively when the marginal buyer is unusually cautious.

4 Applications

We have seen that optimal model training depends on how users will account for imperfections in the AI prediction, and that AI optimally trades off overall accuracy for conditional correctness depending on human parameters. We now consider what optimal training looks like in environments such as adversarial counterparties, competition among model-using agents, committee votes based on model predictions, and so on matter because they transform the same objects already identified in the baseline model. Once those transformed objects are known, the earlier logic tells us how training, reporting, and use should change.

For comparability, in each application we focus on the two objects that drive optimal behavior: the feasible frontier $A(\cdot)$ and the regime-specific iso-payoff slopes

$$\tau := \frac{B - \bar{L}}{H - \bar{L}}, \qquad \sigma := \frac{c_V}{H - B}.$$

Once the transformed objects are identified, the earlier logic applies unchanged. A flatter iso-payoff line shifts the optimum toward higher coverage; a steeper line shifts it toward greater selectivity. An inward shift of the frontier lowers adoption and, when it lowers conditional correctness, makes verification more attractive relative to trust. Each application therefore answers three questions: which object changes, which earlier result now does the work, and whether the environment shifts value creation primarily toward selective training, richer posterior reporting, or organizational changes in who verifies and how.

Table 2 summarizes the mapping from environment to design implication that the applications repeatedly use.

4.1 Adversarial manipulation

Adversarial environments are a natural place to apply the baseline logic because many high-stakes AI deployments involve strategic counterparties who can manipulate inputs, outputs, or downstream interpretation. The question is whether such environments simply call for a more selective model, or whether they also change the relative value of uncertainty reporting and verification. In the language of the baseline model, adversaries are interesting because they can both contract the feasible frontier and move the trust–verify boundary.

Suppose that, conditional on a prediction being issued, with probability $\lambda \in [0, 1)$ an adversary corrupts the recommendation before verification so that the recommendation is incorrect. Alternatively, the adversary with some probability chooses a state which the model cannot predict correctly.

Proposition 12 (Adversarial manipulation)**.** *Let* $p_\lambda(q) := (1-\lambda)p(q)$ *and*

$$A_\lambda(q) := (1-\lambda)A(q).$$

Then:

(i) Trust and verify payoffs become

$$\Pi_T^\lambda(q) = B + A_\lambda(q)(H - \bar{L}) - q(B - \bar{L}) - c,$$

$$\Pi_V^\lambda(q) = B + A_\lambda(q)(H - B) - qc_V - c.$$

Thus adversarial corruption contracts the frontier from A *to* $A_\lambda = (1-\lambda)A$*.*

Table 2: How environment maps into training, reporting, and organizational design

Environment feature	**Model object changed**	**Training implication**	**Reporting or organizational implication**
Adversarial manipulation	Frontier contracts; trust/verify boundary shifts	Favor lower coverage within regime; regime switches may still occur	Emphasize auditability and uncertainty disclosure; assign verification responsibility in strategic settings
Imperfect verification	Verify slope σ steepens; verify region shrinks	Selective training becomes more important	Improve the verification process itself; confidence reporting helps only if it makes review more reliable
O-ring tasks	Frontier deforms to A_M; verify slope rises to σ_M	Favor higher per-step conditional correctness	Localize likely failure points so review can target bottlenecks
Committees	Effective verification cost rises to kc_V	Lower coverage in the verify regime	Primary margin is organizational: designate a pivotal verifying coalition
Payoff externalities across users	Marginal value of correctness changes within regime	Competition favors selectivity; complementarities support broader coverage	Segmentation or customizable thresholds become more valuable
Limited liability	Trust slope τ falls; trust region expands	Private actors choose overly aggressive coverage	Governance, warnings, and oversight matter more than richer scores alone

(ii) Within a fixed regime, the tangency conditions become

$$A'(q_T^\star) = \frac{\tau}{1-\lambda}, \qquad A'(q_V^\star) = \frac{\sigma}{1-\lambda}.$$

Hence corruption steepens both iso-payoff lines and shifts training toward lower coverage.

(iii) The trust-verify comparison becomes

$$\Pi_V^\lambda(q) - \Pi_T^\lambda(q) = q\Big(\big[1-(1-\lambda)p(q)\big](B-\bar{L}) - c_V\Big).$$

So the verify region expands and the region in which trust beats verification shrinks.

(iv) Within a fixed regime, adversarial manipulation lowers the optimal coverage and raises the model's intrinsic conditional correctness $p(q)$. However, the realized *post-corruption correctness $p_\lambda(q) = (1-\lambda)p(q)$ need not increase: the direct degradation from corruption can outweigh the indirect gain from greater selectivity. Separately, because corruption also expands the verify region (part (iii)), a trust-to-verify regime switch can cause the globally optimal coverage to* increase*: an agent who would have trusted a highly selective model may, once corruption lowers conditional correctness, prefer to verify a broader model instead.*

The application maps directly into the baseline objects. Part (i) contracts the frontier; part (ii) then invokes the tangency logic of Proposition 4, implying lower coverage within either regime; part (iii) invokes the trust–verify comparison, implying that verification becomes relatively more attractive. The important point is that adversarial environments need not merely shrink deployment. They can also change the regime in which deployment is optimal.

4.2 Imperfect verification and automation bias

Human-in-the-loop deployment is often justified on the ground that review can correct model mistakes. That makes noisy verification a natural application. If users miss errors or defer too readily to the model, the value of verification falls, and the relevant question becomes whether the right response is better interfaces for review or a more selective model that relies less on review in the first place.

Suppose human verification is noisy. If a prediction is wrong, verification catches the error only with probability $\delta \in (0,1]$. With probability $1-\delta$, the user fails to detect the error and proceeds with the incorrect action. Correct predictions are always verified as correct.

Proposition 13 (Imperfect verification)**.** *Let $\delta \in (0,1]$ denote the probability that verification catches an incorrect recommendation. Then:*

(i) The verify payoff becomes

$$\Pi_V^\delta(q) = B + A(q)\Big(H - \big[\delta B + (1-\delta)\bar{L}\big]\Big) - q\Big(c_V + (1-\delta)(B-\bar{L})\Big) - c.$$

(ii) Trust is unchanged, but the verify-regime iso-payoff slope is replaced by

$$\sigma_\delta := \frac{c_V + (1-\delta)(B-\bar{L})}{H - \big[\delta B + (1-\delta)\bar{L}\big]}.$$

Whenever verification is potentially relevant (in particular, when $c_V < H-B$), lowering δ raises σ_δ and therefore shifts training toward lower coverage.

(iii) The trust-verify comparison becomes

$$\Pi_V^\delta(q) - \Pi_T(q) = q\Big(\delta(1-p(q))(B-\bar{L}) - c_V\Big).$$

Thus trust weakly dominates verify after a prediction iff

$$c_V \geq \delta(1-p(q))(B-\bar{L}).$$

Noisier verification therefore shrinks the verify region.

(iv) Hence automation bias shifts agents away from verify and toward trust or abstain. If verify remains optimal, it calls for a more selective model.

This application leaves the trust side unchanged and worsens only the verify option. Proposition 4 therefore applies through a steeper verify-regime slope, while the trust–verify comparison implies that verification is optimal for a smaller set of environments. The main lesson is not simply that worse reviewers require more accurate models. It is that once review is unreliable, the burden of selection moves back toward the developer.

4.3 O-ring tasks

Many AI deployments are multi-step. Agentic workflows, diagnosis pipelines, and complex coding or planning tasks often require a chain of successful sub-decisions, so that one bad step can destroy task value. That makes O-ring environments a useful application of the baseline model. To keep the multistep problem tractable, however, we study a restricted benchmark rather than the full dynamic control problem.

Suppose the overall task consists of $M \geq 2$ independent subproblems, all of which must be handled correctly to obtain payoff H. If a single subproblem is handled incorrectly, payoff is $\bar{L}$. That is, this is a Kremer (1993) O-ring problem.[8] Impose two additional restrictions. First, if any subproblem abstains, or if any verified subproblem is found to be incorrect, the

[8]Gans and Goldfarb (2026) take a similar perspective on automation.

entire task reverts to the baseline value B; there is no partial salvage. Second, conditional on using the AI, the same continuation rule is applied to every predicted subproblem: either all predicted subproblems are trusted or all predicted subproblems are verified. Richer mixed policies can do better in principle and would require a separate analysis.

Proposition 14 (O-ring tasks under all-or-nothing continuation)**.** *Let the same model be queried independently on each of M subproblems under the all-or-nothing continuation rule just described. Define task-level coverage*

$$Q := q^M$$

and the induced task-level frontier

$$A_M(Q) := A(Q^{1/M})^M.$$

Then task-level conditional correctness is

$$p_M(Q) := \frac{A_M(Q)}{Q} = p(Q^{1/M})^M.$$

(i) Under trust,

$$\Pi_T^{(M)}(Q) = B + A_M(Q)(H - \bar{L}) - Q(B - \bar{L}) - Mc.$$

If all predicted subproblems are verified, then

$$\Pi_V^{(M)}(Q) = B + A_M(Q)(H - B) - Q\,Mc_V - Mc.$$

(ii) In task-level coordinates (Q, A_M), O-ring complementarities leave the trust-regime slope unchanged at

$$\tau = \frac{B - \bar{L}}{H - \bar{L}},$$

but raise the verify-regime slope to

$$\sigma_M = \frac{Mc_V}{H - B}.$$

(iii) Within this restricted policy class, the trust-verify comparison becomes

$$\Pi_V^{(M)}(Q) - \Pi_T^{(M)}(Q) = Q\Big((1 - p_M(Q))(B - \bar{L}) - Mc_V\Big).$$

As M rises, task-level conditional correctness p_M falls, which increases the benefit of

verification, but the verification cost term Mc_V also grows. For moderate M and low c_V, the first effect dominates and the verify region expands; for large M, verification costs can dominate and push the agent toward trust or non-adoption.

(iv) If $c > 0$, then there exists $M^\star$ such that for all $M \geq M^\star$ neither all-trust nor all-verify deployment is adopted.

Within this restricted benchmark, O-ring problems act through both objects at once. They deform the frontier into the task-level frontier A_M and they raise the effective cost of verification by multiplying the relevant coverage cost by M. Proposition 4 therefore applies to the transformed frontier and the steeper verify slope, while the trust–verify comparison determines whether the environment moves the agent toward review, trust, or non-adoption within the restricted policy class. If A_M is not concave, the relevant object for training is its concave hull. A fuller multistep model, in which the agent can verify some steps, solve abstentions separately, or mix AI and human handling across subproblems, is left outside the scope of this proposition.

4.4 Committees and organizational verification

Many consequential decisions are made by boards, panels, and review committees rather than by a single user. In such settings, the existence of a verify option is not enough; verification must also be organized. This makes committees a useful application because they show that the effective value of verification depends not only on c_V but also on how responsibility for checking is assigned.

A committee can implement verification, but doing so requires enough members to verify to overturn a bad recommendation.

Proposition 15 (Committees as effective verification cost)**.** *Consider an odd committee of size $M \geq 3$ that observes a common recommendation.*[9] *After a prediction is issued, members simultaneously choose whether to verify. Verification costs c_V per verifier and perfectly reveals whether the recommendation is correct. A verifier votes to reject if and only if verification shows the recommendation is wrong; a non-verifier votes to accept. The committee follows the recommendation unless a strict majority votes to reject, in which case it reverts to the baseline action. Let*

$$k := \frac{M+1}{2}.$$

[9]The restriction to an odd number of committee members is to simplify tiebreaking concerns.

(i) If exactly k members verify, then the committee implements perfect verification and the committee payoff is

$$\Pi_V^{\text{grp}}(q) = B + A(q)(H - B) - q\,kc_V - c.$$

Thus majority review maps into the baseline model with effective verification cost

$$c_V^{\text{grp}} = kc_V.$$

(ii) The profile in which exactly k designated members verify and the remaining $M - k$ members do not verify is a Nash equilibrium whenever

$$c_V \leq (1 - p)(B - \bar{L}).$$

(iii) The verify-regime slope therefore rises from

$$\sigma = \frac{c_V}{H - B} \quad \textit{to} \quad \sigma_{\text{grp}} = \frac{kc_V}{H - B}.$$

Relative to an individual decision maker, committee verification shifts training toward lower coverage.

(iv) The no-verification profile is also an equilibrium. Hence committees create an organizational implementation problem: without assigned responsibility, behavior can collapse to trust or abstain; with a pivotal verifying coalition, the reduced form is exactly the baseline verify regime with a higher effective verification cost.

The proposition shows that committees matter not because they create a new decision problem, but because they endogenize the verify option. Once a pivotal coalition is formed, Proposition 4 applies with a steeper effective verify slope. Without such a coalition, the system can behave as if verification were absent even when verification is technically feasible.

4.5 Many users and payoff externalities

AI systems are often deployed to many users whose payoffs interact. Traders may compete for the same opportunity; workers may benefit when others also succeed; teams may be complements while firms are rivals. This makes many-user environments interesting because they change the marginal value of correctness without necessarily changing the basic trust–verify boundary.

The value of a correct prediction may depend on how many users of the model also get their tasks right. This nests both complementarity and competition.

Proposition 16 (Many users and payoff externalities)**.** *Suppose $N \geq 2$ symmetric users employ the same model, each on an independent instance, and correctness events are independent across users conditional on q. Let $H(n)$ denote the payoff to a user who is correct when exactly n users in total are correct. Define*

$$h_N(a) := \sum_{k=0}^{N-1} H(1+k) \binom{N-1}{k} a^k (1-a)^{N-1-k},$$

the expected payoff conditional on being correct.

Then:

(i) Trust and verify payoffs for a representative user are

$$\Pi_T^{(N)}(q) = B + a\big[h_N(a) - \bar{L}\big] - q(B - \bar{L}) - c,$$

$$\Pi_V^{(N)}(q) = B + a\big[h_N(a) - B\big] - qc_V - c.$$

(ii) If an optimum is interior within a fixed regime, the local tangency conditions are

$$A'(q_T^\star) = \frac{B - \bar{L}}{h_N(a^\star) + a^\star h_N'(a^\star) - \bar{L}},$$

$$A'(q_V^\star) = \frac{c_V}{h_N(a^\star) + a^\star h_N'(a^\star) - B}.$$

Thus payoff externalities replace the constant correct-action reward H by the endogenous marginal value

$$h_N(a) + a h_N'(a).$$

(iii) If $H(n)$ is decreasing in n (competition), $h_N'(a) \leq 0$, so the local iso-payoff slopes at any interior optimum steepen relative to the single-user case, shifting the local tangency toward lower coverage and greater selectivity. If $H(n)$ is increasing in n (complementarities), $h_N'(a) \geq 0$, and the local slopes flatten, shifting the tangency toward broader coverage. Because h_N depends on $a = A(q)$, the objective need not be globally concave, so these statements characterize interior local optima; the global optimum is obtained by comparing all local candidates with the boundary and baseline values.

(iv) The trust-verify comparison is unchanged:

$$\Pi_V^{(N)}(q) - \Pi_T^{(N)}(q) = q\Big((1 - p(q))(B - \bar{L}) - c_V\Big).$$

Hence many-user externalities mainly change the optimal training target within a regime, not the basic trust-verify boundary.

This application, therefore, uses the main results in a narrower way than the preceding ones. The trust–verify boundary is unchanged, but Proposition 4 now operates with an endogenous marginal value of correctness. Competition steepens the relevant iso-payoff line and favors selectivity; complementarities flatten it and support broader coverage. Because the objective need not be globally concave, the main result should be read as a local characterization of interior candidates rather than a global monotone comparative static.

4.6 Limited liability

Limited liability is a natural application because many model deployers are insulated from the full downside of mistakes through corporate form, insurance, indemnification, or platform governance. This is exactly the kind of environment in which private deployment incentives can diverge from social ones. In the baseline model, that wedge appears through the trust-regime slope.

Proposition 17 (Limited liability)**.** *Suppose that whenever an incorrect action would yield payoff L, the decision maker instead receives*

$$\max\{L, L_{\min}\}.$$

Let

$$\bar{L}_{LL} := \mathbb{E}[\max\{L, L_{\min}\}] \geq \bar{L}.$$

Then:

(i) The trust payoff rises to

$$\Pi_T^{LL}(q) = B + A(q)(H - \bar{L}_{LL}) - q(B - \bar{L}_{LL}) - c.$$

(ii) The trust-regime slope falls from

$$\tau = \frac{B - \bar{L}}{H - \bar{L}} \quad \textit{to} \quad \tau_{LL} := \frac{B - \bar{L}_{LL}}{H - \bar{L}_{LL}} \leq \tau.$$

(iii) The trust-verify boundary becomes

$$c_V \geq (1-p)(B - \bar{L}_{LL}),$$

which is weaker than

$$c_V \geq (1-p)(B - \bar{L}).$$

So the trust region expands.

(iv) Hence limited liability shifts agents toward trust and, when the trust optimum is interior, shifts training toward higher coverage.

The application is a direct use of the baseline trust logic. Limited liability changes neither the frontier nor the verify problem. It lowers the trust-regime slope and expands the trust region, so Proposition 4 implies higher coverage in trust, while the trust–verify comparison implies greater willingness to rely on trust rather than verification. The economic importance of the application is that a familiar governance friction appears through exactly the same object that organizes the baseline theory.

5 Conclusion

Models and predictions, whether in traditional statistics, machine learning, or artificial intelligence, are valuable only through the decision functions that build on them. The organizing question in this paper has therefore been who solves the selection problem created by uncertain model outputs. When users cannot distinguish good predictions from bad ones once the model speaks, the developer must solve that problem mainly at training time by choosing how selective the model should be. When users can cheaply observe calibrated reliability, part of that burden shifts to reporting design and user policy. When users differ, or when the model is supplied by a monopolist, one common training target places that burden on the wrong margin or for the wrong people.

This perspective yields three practical rules. First, when outputs are coarse, miscalibrated, or costly to inspect, model training should target the right coverage–conditional-correctness tradeoff rather than unconditional accuracy. In that environment, refusal behavior, thresholding, and other levers that change selectivity are central design objects. Second, when well-calibrated posteriors can be exposed and cheaply read, developers should invest in posterior informativeness and interface design, because users can then trust, verify, or ignore predictions on a case-by-case basis. Third, when users differ in downside risk or verification cost, a single common model is generally mis-targeted; segmentation, user-specific

thresholds, or richer score reporting can mitigate that loss. The applications in Section 4 follow the same template: adversaries, noisy oversight, O-ring complementarities, committees, competition, and liability matter because they shift either the feasible frontier or the relevant iso-payoff slopes, and small changes in those objects can switch the optimal regime entirely.

Machine learning researchers are not unaware of this problem. Since Chow (1970), the literature has asked how one might train a model when mistaken actions are worse than abstaining. The learning-to-defer literature beginning with Madras et al. (2018) further attempts to learn the policy under which a given decision should be allocated to a human instead of a model on the basis of observed outcomes. In the context of large language models, there exist benchmarks which explicitly punish inaccurate predictions rather than just rank by accuracy, such as AA-Omniscience (Jackson et al., 2025). Our contribution is to begin instead with the parameters of the human utility function and production technology as primitives, derive how humans will use model outputs, and then map those primitives into training design, reporting design, and deployment. That is, *a priori knowledge of the type of human decision functions which a model will be deployed alongside can inform model training in order to maximize human utility when using the model.*

The headline may sound obvious: the best model is the one that creates the most value for its users. But AI is still much more likely to be trained, and benchmarked, to maximize unconditional accuracy than to maximize value. In a fully autonomous setting, or in a Blackwell-orderable setting with easily usable calibrated scores, that can be defensible. Actual use of AI today is rarely autonomous, and rarely well-calibrated enough that users can costlessly recover the right deployment rule from the model's own outputs. Making AI models useful given that humans will make decisions with their predictions is therefore valuable in its own right. It may also matter for the rents humans earn in a world of ubiquitous AI: when models are trained for autonomy rather than for their role as complements to human decision makers, they both limit productivity and distort it toward capital- rather than labor-augmentation, as Acemoglu et al. (2026) among others have pointed out.

6 Proofs

6.1 Proof of Proposition 4

Part (i): on I_T, the objective is

$$\Pi_T(q) = B + A(q)(H - \bar{L}) - q(B - \bar{L}) - c,$$

which is concave because A is concave. If the maximizer is interior, the first-order condition is

$$A'(q_T^\star)(H - \bar{L}) - (B - \bar{L}) = 0,$$

equivalently $A'(q_T^\star) = \tau$. The verify case is identical:

$$\Pi_V(q) = B + A(q)(H - B) - qc_V - c,$$

so an interior maximizer satisfies

$$A'(q_V^\star)(H - B) - c_V = 0,$$

equivalently $A'(q_V^\star) = \sigma$. If the tangency point lies outside the relevant regime interval, concavity implies the interval-constrained maximizer is attained at the relevant boundary point.

Part (ii): if q^a is an interior maximizer of A, then $A'(q^a) = 0$. Coincidence with an interior trust optimum therefore requires $\tau = 0$ (impossible with $B > \bar{L}$), and coincidence with an interior verify optimum requires $\sigma = 0$, i.e. $c_V = 0$. Now suppose q^a is a boundary maximizer, say $q^a = 1$ with $A'(1) > 0$. The regime-specific objective under trust is $\Pi_T(q) = B + A(q)(H - \bar{L}) - q(B - \bar{L}) - c$, which is increasing at $q = 1$ iff $A'(1)(H - \bar{L}) \geq B - \bar{L}$, i.e. $A'(1) \geq \tau$. Likewise the verify objective is increasing at $q = 1$ iff $A'(1) \geq \sigma$. For $q^a = 1$ to be economically optimal, the relevant condition must hold in whichever regime contains $q = 1$. Note that $q = 1$ need not lie in the trust regime at all: when $(1 - p(1))(B - \bar{L}) > c_V$, the point $q = 1$ is in the verify regime and only $A'(1) \geq \sigma$ is required.

Part (iii): $\partial\tau/\partial\bar{L} = -(H - B)/(H - \bar{L})^2 < 0$, so increasing $B - \bar{L}$ raises τ. Since A' is strictly decreasing, the corresponding interior trust optimum moves to lower q. Likewise, increasing c_V raises σ, so the interior verify optimum moves to lower q.

Part (iv): by Lemma 2, each regime corresponds to an interval. By concavity, the maximizer on each regime interval is either the interior tangency from part (i) or a boundary point of that interval. Comparing the resulting finitely many candidates yields the best

adopted model. Final adoption compares that value with the baseline B.

6.2 Proof of Proposition 5

Part (i). On the interval I, each type θ uses a fixed branch. Types in T_I contribute $W_\theta^T(q) = B_\theta + A(q)(H - \bar{L}_\theta) - q(B_\theta - \bar{L}_\theta) - c$, types in V_I contribute $W_\theta^V(q) = B_\theta + A(q)(H - B_\theta) - q\,c_{V,\theta} - c$, and types using the baseline contribute $W_\theta^0(q) = B_\theta$. Integrating,

$$\mathcal{W}(q) = \underbrace{\int B_\theta \, d\mu - \mu(T_I \cup V_I)\, c}_{C_I} + A(q)\left[\int_{T_I} (H - \bar{L}_\theta)\, d\mu + \int_{V_I} (H - B_\theta)\, d\mu\right] - q\left[\int_{T_I} (B_\theta - \bar{L}_\theta)\, d\mu + \int_{V_I} c_{V,\theta}\, d\mu\right].$$

If $\mu(T_I \cup V_I) = 0$, then $\mathcal{W}(q) = \int B_\theta \, d\mu(\theta)$ is constant on I, hence concave.

Part (ii). Differentiating $\mathcal{W}$ on I and setting the derivative to zero:

$$A'(q_I^\dagger)\left[\int_{T_I} (H - \bar{L}_\theta)\, d\mu + \int_{V_I} (H - B_\theta)\, d\mu\right] = \int_{T_I} (B_\theta - \bar{L}_\theta)\, d\mu + \int_{V_I} c_{V,\theta}\, d\mu,$$

which rearranges to (8). When $V_I = \varnothing$, this reduces to $A'(q^\dagger) = \mathbb{E}_{T_I}[B_\theta - \bar{L}_\theta]/\mathbb{E}_{T_I}[H - \bar{L}_\theta]$; when $T_I = \varnothing$, to $A'(q^\dagger) = \mathbb{E}_{V_I}[c_{V,\theta}]/\mathbb{E}_{V_I}[H - B_\theta]$.

Part (iii). With finite support, $\mathcal{W}$ is piecewise concave: each branch-switch creates a new interval. The global optimum is the best among finitely many interval-specific candidates from part (ii) and the boundary points where some type's maximizing branch changes.

Corollary 2. Fix type θ on a branch $b \in \{T, V\}$ that is constant between $q_\theta^\star$ and $q^\dagger$. The branch-specific welfare function is $W_{\theta,b}(q) = B_\theta + A(q)\alpha_b - q\beta_b - c$ for positive constants α_b, β_b determined by the branch and agent type. Since A is twice differentiable on the interval and strictly concave, $W_{\theta,b}$ is twice differentiable and strictly concave with $W'_{\theta,b}(q_\theta^\star) = 0$. Taylor's theorem with Lagrange remainder gives

$$W_{\theta,b}(q^\dagger) = W_{\theta,b}(q_\theta^\star) + \tfrac{1}{2}\, W''_{\theta,b}(\tilde{q}_\theta)\,(q^\dagger - q_\theta^\star)^2$$

for some $\tilde{q}_\theta$ between $q^\dagger$ and $q_\theta^\star$. Since $W''_{\theta,b} = A'' \cdot \alpha_b < 0$, rearranging gives (9). If the maximizing branch changes between $q_\theta^\star$ and $q^\dagger$, then $\Delta_\theta \geq 0$ holds directly because $q_\theta^\star$ maximizes W_θ by definition, but the exact quadratic representation need not hold since the branch-specific second-order expansion no longer applies across the full interval. □

6.3 Proof of Proposition 9

Part (i). By Proposition 8(iii), the optimal prediction set at any coverage level q is the upper tail of the posterior distribution. Under trust-only deployment on the upper tail, the objective is

$$\Pi_T(q) = B + \mathcal{A}(q)(H - \bar{L}) - q(B - \bar{L}) - c,$$

which is concave because $\mathcal{A}$ is concave. The first-order condition for an interior optimum is

$$\mathcal{A}'(q_T^\star)(H - \bar{L}) - (B - \bar{L}) = 0,$$

equivalently $\mathcal{A}'(q_T^\star) = \tau$. For verify-only deployment,

$$\Pi_V(q) = B + \mathcal{A}(q)(H - B) - q\, c_V - c,$$

with interior first-order condition $\mathcal{A}'(q_V^\star)(H - B) - c_V = 0$, giving $\mathcal{A}'(q_V^\star) = \sigma$.

Part (ii). Under threshold deployment $\{\pi \geq t\}$ with pointwise-optimal continuation, the developer's value is

$$V(t) = B + \int_t^{\bar{\pi}} \phi(\pi)\, f(\pi)\, d\pi - c.$$

By the fundamental theorem of calculus, $V'(t) = -\phi(t)f(t)$. Since $f > 0$ on $[\underline{\pi}, \bar{\pi}]$, the sign of $V'(t)$ is determined by $\phi(t)$.

We verify that $\phi(\pi) = 0$ for $\pi \leq s^* := \min\{\sigma, \tau\}$ and $\phi(\pi) > 0$ for $\pi > s^*$.

Case (a): $c_V < (B-\bar{L})(H-B)/(H-\bar{L})$. Then $\sigma < \tau$, so $s^* = \sigma$. For $\pi < \sigma$: $\pi(H-B) < c_V$ and $\pi(H - \bar{L}) < \sigma(H - \bar{L}) = c_V(H - \bar{L})/(H - B) < B - \bar{L}$, where the last inequality uses $c_V < (B - \bar{L})(H - B)/(H - \bar{L})$. Hence both terms in ϕ are negative, so $\phi(\pi) = 0$. For π slightly above σ: $\pi(H - B) > c_V$, so $\phi(\pi) \geq \pi(H - B) - c_V > 0$.

Case (b): $c_V \geq (B-\bar{L})(H-B)/(H-\bar{L})$. Then $\sigma \geq \tau$, so $s^* = \tau$. For $\pi < \tau$: $\pi(H-\bar{L}) < B-\bar{L}$ and $\pi(H - B) < \tau(H - B) = (B - \bar{L})(H - B)/(H - \bar{L}) \leq c_V$, so $\phi(\pi) = 0$. For $\pi > \tau$: $\pi(H - \bar{L}) > B - \bar{L}$, so $\phi(\pi) \geq \pi(H - \bar{L}) - (B - \bar{L}) > 0$.

In both cases, V is constant for all feasible $t \leq s^\star$ and strictly decreasing for $t > s^\star$. Hence every feasible $t \leq s^\star$ achieves the global maximum, and $t^\star = s^\star$ is the most selective optimal threshold.

6.4 Proof of Proposition 11

We proceed part by part.

Part (i). By assumption, a type-s buyer obtains gross surplus $u(s,q) = A(q) - sq$. If the model is made available at zero price to all buyers, aggregate welfare is

$$\mathcal{W}(q) = \int_{\underline{s}}^{\bar{s}} \big[A(q) - sq\big]\, dG(s) = A(q) - q \int_{\underline{s}}^{\bar{s}} s\, dG(s) = A(q) - q\,\mathbb{E}[s].$$

Since A is strictly concave, $\mathcal{W}$ is strictly concave. If $A'(1) < \mathbb{E}[s]$, the unique interior optimum satisfies $\mathcal{W}'(q) = A'(q) - \mathbb{E}[s] = 0$, giving (18). If $A'(1) \geq \mathbb{E}[s]$, then $\mathcal{W}'(q) > 0$ for all $q \in [0,1)$ and $q^S = 1$.

Part (ii). Fix (q,r). A buyer of type s purchases if and only if

$$A(q) - sq \geq r.$$

Since the left-hand side is strictly decreasing in s, the set of buyers is an interval of the form $[\underline{s}, \hat{s}]$ for some cutoff type $\hat{s}$. Rearranging the indifference condition for the cutoff buyer gives

$$r = A(q) - \hat{s}q.$$

Hence demand is $G(\hat{s})$ and profit is

$$\Pi(q,\hat{s}) = r\, G(\hat{s}) = \big[A(q) - \hat{s}q\big]\, G(\hat{s}),$$

which is (19).

Part (iii). Consider an interior optimum $(q^M, \hat{s}^M)$ of (19). Differentiating with respect to q gives

$$\frac{\partial \Pi}{\partial q} = \big[A'(q) - \hat{s}\big]\, G(\hat{s}).$$

At an interior optimum, $G(\hat{s}^M) > 0$, so the first-order condition is

$$A'(q^M) = \hat{s}^M,$$

which is (20).

Differentiating (19) with respect to $\hat{s}$ gives

$$\frac{\partial \Pi}{\partial \hat{s}} = -q\, G(\hat{s}) + \big[A(q) - \hat{s}q\big]\, g(\hat{s}).$$

Setting this equal to zero at an interior optimum yields

$$A(q^M) - \hat{s}^M q^M = q^M \, \frac{G(\hat{s}^M)}{g(\hat{s}^M)},$$

which is (21).

Part (iv). If both q^S and q^M are interior, then from part (i) the planner's choice satisfies $A'(q^S) = \mathbb{E}[s]$, and from part (iii) the monopolist's choice satisfies $A'(q^M) = \hat{s}^M$. Because A' is strictly decreasing,

$$q^M > q^S \iff A'(q^M) < A'(q^S) \iff \hat{s}^M < \mathbb{E}[s].$$

Likewise, $q^M < q^S \iff \hat{s}^M > \mathbb{E}[s]$. This proves (22). When one or both solutions are at $q = 1$, the comparison must be made directly between the boundary and interior candidates.

6.5 Proof of Proposition 12

Conditional on a prediction being issued, corruption makes the recommendation correct with probability

$$p_\lambda(q) = (1 - \lambda)p(q),$$

so unconditional accuracy becomes

$$A_\lambda(q) = q \, p_\lambda(q) = (1 - \lambda)A(q).$$

Under trust,

$$\Pi_T^\lambda(q) = q\Big[p_\lambda(q)H + \big(1 - p_\lambda(q)\big)\bar{L}\Big] + (1 - q)B - c = B + A_\lambda(q)(H - \bar{L}) - q(B - \bar{L}) - c.$$

Under verify, any incorrect recommendation—native or corrupted—is replaced by the baseline after verification, so

$$\Pi_V^\lambda(q) = q\Big[p_\lambda(q)H + \big(1 - p_\lambda(q)\big)B - c_V\Big] + (1 - q)B - c = B + A_\lambda(q)(H - B) - qc_V - c.$$

Within either regime, the objective is the baseline objective with A replaced by $(1-\lambda)A$. Therefore an interior trust optimum satisfies

$$(1 - \lambda)A'(q)(H - \bar{L}) - (B - \bar{L}) = 0,$$

hence

$$A'(q_T^\star) = \frac{\tau}{1-\lambda},$$

and similarly

$$A'(q_V^\star) = \frac{\sigma}{1-\lambda}.$$

Finally,

$$\Pi_V^\lambda(q) - \Pi_T^\lambda(q) = q\Big(\big[1 - p_\lambda(q)\big](B - \bar{L}) - c_V\Big) = q\Big(\big[1 - (1-\lambda)p(q)\big](B - \bar{L}) - c_V\Big).$$

Since $1 - (1-\lambda)p(q) > 1 - p(q)$ for every $\lambda > 0$, the verify region expands and the trust region shrinks.

Parts (i)–(iii) establish the within-regime comparative statics. Within trust, the tangency condition $A'(q_T^\star) = \tau/(1-\lambda)$ and the fact that A' is strictly decreasing imply that $q_T^\star$ falls as λ rises. Within verify, the same logic applies to $A'(q_V^\star) = \sigma/(1-\lambda)$. However, the global optimum is $\max\{\Pi_T^\lambda(q_T^\star),\, \Pi_V^\lambda(q_V^\star),\, B\}$, and since part (iii) shows that corruption expands the verify region, the global optimum can switch from trust to verify as λ increases. Because $q_V^\star > q_T^\star$ is possible at the new parameter values, the globally optimal coverage need not decrease.

6.6 Proof of Proposition 13

If a prediction is verified, then with probability $p(q)$ it is correct and yields H. With probability $1 - p(q)$ it is wrong; verification catches the error with probability δ, yielding the baseline B, and misses it with probability $1 - \delta$, yielding $\bar{L}$. Therefore

$$\Pi_V^\delta(q) = q\Big[p(q)H + \big(1 - p(q)\big)\big(\delta B + (1-\delta)\bar{L}\big) - c_V\Big] + (1-q)B - c.$$

Using $A(q) = qp(q)$ and $q(1 - p(q)) = q - A(q)$,

$$\Pi_V^\delta(q) = B + A(q)\Big(H - \big[\delta B + (1-\delta)\bar{L}\big]\Big) - q\Big(c_V + (1-\delta)(B - \bar{L})\Big) - c.$$

This proves part (i).

Part (ii) follows because the verify problem is now linear in $A(q)$ and q with slope

$$\sigma_\delta = \frac{c_V + (1-\delta)(B - \bar{L})}{H - \big[\delta B + (1-\delta)\bar{L}\big]}.$$

Write $X := B - \bar{L} > 0$ and $Y := H - B > 0$. Then

$$\sigma_\delta = \frac{c_V + (1-\delta)X}{Y + (1-\delta)X},$$

so

$$\frac{\partial \sigma_\delta}{\partial \delta} = \frac{X(c_V - Y)}{(Y + (1-\delta)X)^2}.$$

Hence, whenever $c_V < Y = H - B$, lowering δ raises σ_δ.

For part (iii),

$$\Pi_V^\delta(q) - \Pi_T(q) = q\Big[(1-p(q))\big(\delta B + (1-\delta)\bar{L} - \bar{L}\big) - c_V\Big] = q\Big(\delta(1-p(q))(B-\bar{L}) - c_V\Big).$$

Thus trust weakly dominates verify after a prediction iff

$$c_V \geq \delta(1-p(q))(B-\bar{L}).$$

Since the right-hand side falls with δ, noisier verification shrinks the verify region.

Part (iv) is then immediate from parts (ii) and (iii).

6.7 Proof of Proposition 14

Under the all-or-nothing restrictions of Proposition 14, if the same model is used on each of the M independent subproblems, then all M subproblems are predicted with probability

$$Q = q^M.$$

All M are predicted and correct with probability

$$A(q)^M.$$

Hence the induced task-level frontier is

$$A_M(Q) = A(Q^{1/M})^M,$$

and task-level conditional correctness is

$$p_M(Q) = \frac{A_M(Q)}{Q} = p(Q^{1/M})^M.$$

Under trust, the overall task yields H if all M are predicted and correct, $\bar{L}$ if all M

are predicted but at least one is wrong, and B if at least one subproblem is not predicted. Therefore

$$\Pi_T^{(M)}(q) = A(q)^M H + \big(q^M - A(q)^M\big)\bar{L} + (1 - q^M)B - Mc,$$

which rearranges to

$$\Pi_T^{(M)}(Q) = B + A_M(Q)(H - \bar{L}) - Q(B - \bar{L}) - Mc.$$

If all predicted subproblems are verified, then any error anywhere triggers reversion to the baseline, so

$$\Pi_V^{(M)}(q) = A(q)^M H + \big(q^M - A(q)^M\big)B - q^M Mc_V - Mc,$$

which rearranges to

$$\Pi_V^{(M)}(Q) = B + A_M(Q)(H - B) - Q\,Mc_V - Mc.$$

This proves part (i).

Part (ii) now follows directly: in task-level coordinates (Q, A_M), the trust objective has slope

$$\frac{B - \bar{L}}{H - \bar{L}} = \tau,$$

while the verify objective has slope

$$\frac{Mc_V}{H - B} = \sigma_M.$$

For part (iii),

$$\Pi_V^{(M)}(Q) - \Pi_T^{(M)}(Q) = Q\Big((1 - p_M(Q))(B - \bar{L}) - Mc_V\Big).$$

Thus, within the restricted policy class, trust becomes less attractive relative to verification as M rises through the decline in $p_M(Q)$, although the linear growth in verification cost can dominate for sufficiently large M.

For part (iv), define $\alpha := \max_{q\in[0,1]} A(q)$. Under the standing assumptions, $A(q) < q$ for all $q > 0$ (since $p(q) = A(q)/q$ is strictly decreasing with $p(0^+) = A'(0) \le 1$), so $\alpha < 1$. Now $A_M(Q) = A(Q^{1/M})^M \le \alpha^M$ for all $Q \in [0, 1]$. Under trust,

$$\Pi_T^{(M)}(Q) - B = A_M(Q)(H - \bar{L}) - Q(B - \bar{L}) - Mc \le \alpha^M(H - \bar{L}) - Mc.$$

Under verify,

$$\Pi_V^{(M)}(Q) - B = A_M(Q)(H - B) - Q\,Mc_V - Mc \leq \alpha^M(H - B) - Mc.$$

Since $\alpha < 1$ and $c > 0$, both upper bounds tend to $-\infty$ as $M \to \infty$. Hence for sufficiently large M neither restricted regime is adopted.

6.8 Proof of Proposition 15

Let $k = (M + 1)/2$. Suppose exactly k members verify and the remaining $M - k$ do not.

If the recommendation is correct, all verifiers observe this and vote to accept, while all non-verifiers also vote to accept. The committee accepts. If the recommendation is incorrect, all verifiers observe this and vote to reject, while all non-verifiers vote to accept. Since

$$k = \frac{M+1}{2} > M - k,$$

the verifiers form a strict majority and the committee rejects, reverting to the baseline. Thus the committee implements perfect verification. Expected payoff is therefore

$$\Pi_V^{\text{grp}}(q) = q\Big[p(q)H + \big(1 - p(q)\big)B - kc_V\Big] + (1 - q)B - c = B + A(q)(H - B) - q\,kc_V - c.$$

This proves part (i).

For part (ii), consider a non-verifier. Deviating to verify does not change the committee decision in either state, because the verifying coalition already forms a strict majority when the recommendation is wrong and the committee accepts when it is correct. The deviation therefore only adds cost c_V, so it is never profitable.

Now consider one of the k verifiers. If this member deviates to not verifying, then when the recommendation is wrong the number of rejecting votes falls from k to $k - 1$, so the committee no longer rejects: the wrong recommendation is implemented. The deviator saves c_V but loses

$$(1 - p)(B - \bar{L})$$

in expected payoff conditional on a prediction. Therefore the deviation is unprofitable iff

$$c_V \leq (1 - p)(B - \bar{L}).$$

For part (iii), the committee verify payoff is exactly the baseline verify payoff with effective

verification cost

$$c_V^{\text{grp}} = kc_V,$$

so the verify-regime slope becomes

$$\sigma_{\text{grp}} = \frac{kc_V}{H - B}.$$

For part (iv), if no one verifies, a unilateral deviation to verifying cannot overturn the committee outcome when the recommendation is wrong, because a single reject vote is not a majority. The deviation only adds cost c_V, so the no-verification profile is also an equilibrium.

6.9 Proof of Proposition 16

Let

$$a := A(q)$$

denote the probability that a given user is correct. Conditional on one user's being correct, the number of other correct users is binomial:

$$K \sim \text{Bin}(N-1, a).$$

Therefore the expected payoff conditional on being correct is

$$h_N(a) = \mathbb{E}[H(1+K)] = \sum_{k=0}^{N-1} H(1+k)\binom{N-1}{k} a^k (1-a)^{N-1-k}.$$

Under trust, a representative user is correct with probability a, is incorrect after a prediction with probability $q - a$, and faces the baseline with probability $1 - q$. Hence

$$\Pi_T^{(N)}(q) = a\, h_N(a) + (q-a)\bar{L} + (1-q)B - c = B + a\big[h_N(a) - \bar{L}\big] - q(B - \bar{L}) - c.$$

Under verify, incorrect predictions are replaced by the baseline, so

$$\Pi_V^{(N)}(q) = a\, h_N(a) + (q-a)B - qc_V + (1-q)B - c = B + a\big[h_N(a) - B\big] - qc_V - c.$$

This proves part (i).

For part (ii), differentiate with respect to q. Since $a = A(q)$,

$$\frac{d}{dq}\Big[a\, h_N(a)\Big] = A'(q)\Big(h_N(a) + ah_N'(a)\Big).$$

Therefore

$$\frac{d\Pi_T^{(N)}}{dq} = A'(q)\Big(h_N(a) + ah'_N(a) - \bar{L}\Big) - (B - \bar{L}),$$

and an interior trust optimum satisfies

$$A'(q_T^\star) = \frac{B - \bar{L}}{h_N(a^\star) + a^\star h'_N(a^\star) - \bar{L}}.$$

Similarly,

$$\frac{d\Pi_V^{(N)}}{dq} = A'(q)\Big(h_N(a) + ah'_N(a) - B\Big) - c_V,$$

so an interior verify optimum satisfies

$$A'(q_V^\star) = \frac{c_V}{h_N(a^\star) + a^\star h'_N(a^\star) - B}.$$

For part (iii), if $H(n)$ is increasing in n, then the binomial distribution of K shifts upward in the sense of first-order stochastic dominance as a rises, so $h_N(a)$ is nondecreasing and $h'_N(a) \geq 0$ wherever the derivative exists. If $H(n)$ is decreasing, the reverse holds and $h'_N(a) \leq 0$. The formulas in part (ii) then imply flatter local slopes in the complementary case and steeper local slopes in the competitive case.

For part (iv),

$$\Pi_V^{(N)}(q) - \Pi_T^{(N)}(q) = a(\bar{L} - B) + q(B - \bar{L}) - qc_V = q\Big((1 - p(q))(B - \bar{L}) - c_V\Big),$$

because $a = qp(q)$. Thus the trust-verify boundary is unchanged.

6.10 Proof of Proposition 17

Pointwise truncation raises the payoff from incorrect actions, so

$$\bar{L}_{LL} \geq \bar{L}.$$

Substituting $\bar{L}_{LL}$ into the trust payoff gives

$$\Pi_T^{LL}(q) = B + A(q)(H - \bar{L}_{LL}) - q(B - \bar{L}_{LL}) - c,$$

proving part (i).

For part (ii), define

$$\tau(\ell) := \frac{B - \ell}{H - \ell}.$$

Then

$$\frac{d\tau}{d\ell} = -\frac{H - B}{(H - \ell)^2} < 0.$$

Since $\bar{L}_{LL} \geq \bar{L}$, it follows that

$$\tau_{LL} = \tau(\bar{L}_{LL}) \leq \tau(\bar{L}) = \tau.$$

For part (iii), the trust-verify comparison in the limited-liability environment is

$$\Pi_V(q) - \Pi_T^{LL}(q) = q\Big((1 - p)(B - \bar{L}_{LL}) - c_V\Big).$$

Since $\bar{L}_{LL} \geq \bar{L}$, one has

$$B - \bar{L}_{LL} \leq B - \bar{L},$$

so the trust condition

$$c_V \geq (1 - p)(B - \bar{L}_{LL})$$

is weaker than the baseline trust condition.

Part (iv) follows immediately: the trust-region slope falls, the trust region expands, and any interior trust optimum moves toward higher coverage by Proposition 4(iii).

References

Acemoglu, Daron, Dingwen Kong, and Asuman Ozdaglar, "AI, Human Cognition and Knowledge Collapse," Working Paper 34910, National Bureau of Economic Research March 2026.

Agarwal, Nikhil, Alex Moehring, and Alexander Wolitzky, "Designing Human-AI Collaboration: A Sufficient-Statistic Approach," Working Paper 33949, National Bureau of Economic Research June 2025.

Agrawal, Ajay, Joshua S. Gans, and Avi Goldfarb, *Prediction Machines: The Simple Economics of Artificial Intelligence*, Boston, MA: Harvard Business Review Press, 2018.

_ , _ , and _ , "The Economics of Bicycles for the Mind," Working paper 2025.

Agrawal, Ajay K, John McHale, and Alexander Oettl, "Enhancing Worker Productivity Without Automating Tasks: A Different Approach to AI and the Task-Based Model," Technical Report, National Bureau of Economic Research 2026.

Athey, Susan C., Kevin A. Bryan, and Joshua S. Gans, "The Allocation of Decision Authority to Human and Artificial Intelligence," *AEA Papers and Proceedings*, 2020, *110*, 80–84.

Autor, David, Andrew Caplin, Daniel J Martin, and Philip Marx, "Misaligned by Design: Incentive Failures in Machine Learning," Technical Report, National Bureau of Economic Research 2025.

Bartlett, Peter L. and Marten H. Wegkamp, "Classification with a Reject Option using a Hinge Loss," *Journal of Machine Learning Research*, 2008, *9*, 1823–1840.

Blackwell, David, "Comparison of Experiments," in Jerzy Neyman, ed., *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley and Los Angeles: University of California Press, 1951, pp. 93–102.

_ , "Equivalent Comparisons of Experiments," *The Annals of Mathematical Statistics*, 1953, *24* (2), 265–272.

Brynjolfsson, Erik, Danielle Li, and Lindsey R. Raymond, "Generative AI at Work," *The Quarterly Journal of Economics*, 2025, *140* (2), 889–942.

Catalini, Christian, Xiang Hui, and Jane Wu, "Some Simple Economics of AGI," *arXiv preprint arXiv:2602.20946*, 2026.

Chow, C. K., "On Optimum Recognition Error and Reject Tradeoff," *IEEE Transactions on Information Theory*, 1970, *16* (1), 41–46.

Cortes, Corinna, Giulia DeSalvo, and Mehryar Mohri, "Learning with Rejection," in "Proceedings of the 27th International Conference on Algorithmic Learning Theory (ALT)" Springer 2016, pp. 67–82.

El-Yaniv, Ran and Yair Wiener, "On the Foundations of Noise-free Selective Classification," *Journal of Machine Learning Research*, 2010, *11* (53), 1605–1641.

Gans, Joshua S., *The Microeconomics of Artificial Intelligence*, MIT Press, 2025.

_ , "A Model of Artificial Jagged Intelligence," Technical Report, National Bureau of Economic Research 2026.

_ , "Optimal Use of Preferences in Artificial Intelligence Algorithms," Technical Report, National Bureau of Economic Research 2026.

_ **and Avi Goldfarb**, "O-Ring Automation," Technical Report, National Bureau of Economic Research 2026.

Geifman, Yonatan and Ran El-Yaniv, "Selective Classification for Deep Neural Networks," in "Advances in Neural Information Processing Systems 30" 2017, pp. 4878–4887.

Guo, Chuan, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger, "On Calibration of Modern Neural Networks," in "Proceedings of the 34th International Conference on Machine Learning (ICML)" 2017, pp. 1321–1330.

Herbei, Radu and Marten Wegkamp, "Classification with Reject Option," *The Canadian Journal of Statistics*, 2006, *34* (4), 709–721.

Jackson, Declan, William Keating, George Cameron, and Micah Hill-Smith, "AA-Omniscience: Evaluating Cross-Domain Knowledge Reliability in Large Language Models," 2025.

Jones, Erik, Shiori Sagawa, Pang Wei Koh, Ananya Kumar, and Percy Liang, "Selective Classification Can Magnify Disparities Across Groups," in "International Conference on Learning Representations" 2021.

Kalai, Adam Tauman, Ofir Nachum, Santosh S. Vempala, and Edwin Zhang, "Why Language Models Hallucinate," 2025. arXiv preprint arXiv:2509.04664.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan, "Human Decisions and Machine Predictions*," *The Quarterly Journal of Economics*, 08 2017, *133* (1), 237–293.

Kremer, Michael, "The O-Ring Theory of Economic Development," *The Quarterly Journal of Economics*, 1993, *108* (3), 551–575.

Madras, David, Toniann Pitassi, and Richard Zemel, "Predict Responsibly: Improving Fairness and Accuracy by Learning to Defer," in "Advances in Neural Information Processing Systems 31 (NeurIPS 2018)" 2018.

Mozannar, Hussein and David Sontag, "Consistent estimators for learning to defer to an expert," in "International Conference on Machine Learning" PMLR 2020, pp. 7076–7087.

Noy, Shakked and Whitney Zhang, "Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence," *Science*, 2023, *381* (6654), 187–192.

Vaccaro, Michelle, Abdullah Almaatouq, and Thomas Malone, "When Combinations of Humans and AI are Useful: A Systematic Review and Meta-analysis," *Nature Human Behaviour*, 2024, *8* (12), 2293–2303.

Wald, Abraham, *Statistical Decision Functions*, New York: John Wiley & Sons, 1950.