WORKING PAPERS

N° 1715

February 2026

"Reputational Conservatism in Expert Advice"

Georgy Lukyanov and Anna Vlasova

Toulouse School of Economics

Reputational Conservatism in Expert Advice[*]

Georgy Lukyanov[†] Anna Vlasova[‡]

Abstract

We develop a tractable career–concerns model of expert recommendations with a continuous private signal. In equilibrium, advice obeys a cutoff rule: the expert recommends the risky option if and only if the signal exceeds a threshold. Under a mild relative–diagnosticity condition, the threshold is (weakly) increasing in reputation, yielding *reputational conservatism.* Signal informativeness and success priors lower the cutoff, while stronger career concerns raise it. A success–contingent bonus implements any target experimentation rate via a one–to–one mapping, providing an implementable design lever.

Keywords: career concerns; expert advice; reputational incentives; information design; experimentation.
JEL: D82; D83; C72.

1 Introduction

Committees, boards, and regulators routinely rely on expert recommendations when the underlying evidence is noisy and outcomes are publicly observed. In such environments, experts care not only about the contemporaneous payoff from getting things right, but also about how their actions and realized outcomes shift the market's assessment of their ability. This creates a classic *career–concerns* tradeoff that practitioners often describe as "playing it safe at the top": senior analysts, established doctors, or reputable policymakers appear more conservative than their lower–reputation counterparts. When does reputation make experts conservative, and what simple levers let a principal raise or lower experimentation in a disciplined way?

[*] We thank Vitalijs Jascisens, Steven Kivinen, Alessandro Riboni, Stepan Svistunov, Alexey Verenikin and Maria Ziskelevich for useful comments and discussions. All remaining errors are ours.

[†] Toulouse School of Economics. Email: georgy.lukyanov@tse-fr.eu.

[‡] CREST—École Polytechnique. Email: anna.vlasova@polytechnique.edu.

We answer these questions in a tractable model of expert advice with a *continuous private signal* about a binary payoff.[1] An expert publicly recommends either a risky or a safe action; risky recommendations are implemented with a probability that (plausibly) increases in the expert's current reputation, and realized successes/failures are publicly observed.[2] The key primitives are minimal: (i) a monotone signal about success, (ii) a reduced–form reputational payoff that depends on the *posterior* after public observations, and (iii) an implementation intensity that is higher for more reputable experts. We deliberately focus on a single episode to isolate the selection effect of reputation on advice, abstracting from dynamic contracting or long-horizon reputation management.

Our first result is a simple characterization: *equilibrium advice follows a cutoff rule.* The expert recommends the risky action if and only if the signal exceeds a threshold that depends on reputation (Proposition 4.1). The cutoff logic rests solely on single–crossing of expected payoffs in the continuous signal and yields uniqueness and continuity of the threshold.

Our central insight is *reputational conservatism*: under a mild relative–diagnosticity condition—failures are (weakly) more informative than successes at higher standing—the cutoff is (weakly) increasing in reputation (Proposition 4.2).[3] Intuitively, when implementation is more likely at high reputation and failures are particularly revealing in that region, high–reputation experts internalize a larger reputational downside from visible failures and thus require stronger evidence before endorsing risk. The model therefore predicts fewer risky recommendations but higher conditional hit rates among high–reputation experts.

We then deliver transparent comparative statics (Proposition 4.3). Greater signal informativeness or a higher success prior *lowers* the cutoff; stronger career–concern intensity *raises* it in the conservatism region. Finally, we provide a design lever with immediate implementation value: a *success–only bonus* induces a one–to–one mapping to the experimentation rate at a given reputation (Proposition 5.1). This gives principals a single–parameter knob to target experimentation without estimating the entire structure. A complementary corollary shows how *gatekeeping/monitoring* that lowers implementation intensity shifts selection on risk in predictable ways (Corollary 5.2).

We offer a clean, continuous–signal foundation for reputational conservatism in expert

[1] The signal satisfies single-crossing (MLRP), not necessarily Normality. Our results rely on monotone likelihood ratios, so they extend to standard continuous families. A Gaussian microfoundation appears in Appendix B.

[2] We treat implementation intensity as a reduced-form object capturing client uptake, committee approval, or regulatory gatekeeping. It is exogenous in the baseline for clarity; policy shifts (e.g., stricter gatekeeping) are modeled by a parameter T that monotonically lowers implementation intensity.

[3] Our RD condition is *local* (evaluated at the cutoff) and weaker than global "bad-news" dominance. Intuitively, near the threshold, failures are (weakly) more revealing than successes for high-reputation experts, especially when their recommendations are implemented more often. Appendix B gives a Gaussian illustration. If RD were violated, the reputational comparative static could be non-monotone.

advice, distinct from classic herding and multi–expert cheap-talk channels. We show that simple success-contingent pay enables direct implementation of target experimentation, providing a design tool orthogonal to full contract redesign. Our theory produces portable, testable predictions about how standing correlates with the frequency and accuracy of risky recommendations, while remaining agnostic about measurement.[4]

2 Related Literature

Our work is closest to models where observable outcomes feed a market's assessment of ability and shape effort or risk choices. Recent contributions include Halac and Kremer (2020), who study dynamic experimentation under career concerns with *bad–news* learning and show how reputational incentives distort quitting; Marinovic and Szydlowski (2022), who analyze continuous–time monitoring with career concerns on the *monitor's* side; and Azrieli (2021), who characterize contracts that induce experts to collect and truthfully report information under monitoring frictions. Relatedly, Moroni (2022) examines experimentation assignments in organizations; and work on delegated experimentation and dynamic delegation (e.g., Escobar and Zhang, 2021; Okat and Nash, 2024; Wittbrodt and Yoder, 2025) studies when principals cede experimentation to agents. Relative to this strand, we isolate a single–episode recommendation problem with continuous signals and derive a global threshold that *increases* in reputation under a mild diagnosticity condition, yielding sharp comparative statics and implementable levers.

The classic view models experts who care about how their messages or recommendations affect reputation (e.g., Ottaviani and P. N. Sørensen, 2006; Ottaviani and P. Sørensen, 2006). Closer to us, Liu and Sanyal (2012) studies expert recommendations under career concerns when actions can be reversed by a principal, and Klein and Mylovanov (2017) analyzes when long horizons counteract conservative reporting incentives. Catonini and Stepanov (2023) studies information aggregation when a reputation–concerned decision maker solicits advice from reputation–concerned advisors.

We also connect to persuasion and information design, particularly results that deliver tractable, global characterizations and implementable instruments. Foundationally, Kamenica and Gentzkow (2011) shows the value of commitment over signal structures. Subsequent advances include persuasion with publicly or privately informed receivers (Kolotilin, 2015; Kolotilin et al., 2017) and robust persuasion under model uncertainty (Dworczak and Pavan,

[4] We deliberately present a theory-only contribution. The predictions (selection on risk by reputation; conditional hit rates; policy invariants for bonuses and gatekeeping) are qualitative and portable across domains, but we do not undertake measurement or calibration here.

2022). Our contribution is complementary: rather than designing a signal, we ask how a principal can shape *recommendations* through simple, success–contingent bonuses and implementation intensity, taking the information source as given.

A growing literature studies how organizational rules and monitoring shape information production or selection on risk (e.g., Marinovic and Szydlowski, 2022; Azrieli, 2021). Our model provides a compact link between gatekeeping (lower implementation intensity for risky recommendations) and selection on risk via the cutoff, and delivers a closed–form mapping from bonuses to experimentation frequencies, which can be used inside committees or review processes.

Across these strands, we emphasize three departures. (i) A continuous-signal environment that yields a unique cutoff and *global* monotone comparative statics without case splits; (ii) a reputational–diagnosticity condition that is microfounded in standard Gaussian settings (Appendix B) and delivers reputational conservatism; (iii) implementable levers—success–only bonuses and gatekeeping—that provide direct, testable mappings to experimentation and selection on risk.

Roadmap. Section 3 presents the environment and equilibrium notion. Section 4 establishes the cutoff characterization, reputational conservatism, and comparative statics. Section 5 develops the implementability of experimentation via success–only bonuses and the effects of gatekeeping. Section 6 discusses implications and scope; the Appendix contains proofs and a Gaussian microfoundation for the diagnosticity condition.

3 Model

3.1 Environment

We study a single recommendation episode. An expert privately observes a continuous signal $x \in \mathbb{R}$ about a binary payoff ("success" $y = 1$ or "failure" $y = 0$). After observing x, the expert recommends either a risky action $a = R$ or a safe action $a = S$.[5]

A public *reputation* state $\rho \in (0, 1)$ summarizes the market's belief about the expert's ability at the start of the episode. If the risky action is recommended, it is *implemented* with probability $\lambda(\rho) \in [0, 1]$; if implemented, the realized outcome $y \in \{0, 1\}$ is observed publicly. If the safe action is recommended, a conservative alternative is taken and produces no informative outcome in that period.

[5] What matters is that the risky option generates an informative, publicly observed outcome when implemented. The safe option can be interpreted as a fallback whose outcome is either uninformative or not publicly salient; allowing a low-informativeness "safe outcome" leaves our results intact.

Conditional on the signal, the risky action would succeed with probability

$$p(x) \in (0,1), \qquad p'(x) > 0,$$

so that higher signals make success more likely (a canonical microfoundation is an MLRP signal about a binary state).

3.2 Timing and beliefs

The within-episode timing is:

(i) Public reputation ρ is given.

(ii) The expert privately observes x.

(iii) The expert recommends $a \in \{R, S\}$; the recommendation is public.

(iv) If $a = R$, the recommendation is implemented with probability $\lambda(\rho)$; if implemented, the outcome $y \in \{0,1\}$ is realized with $\mathbb{P}[y = 1 \mid x] = p(x)$ and publicly observed.

(v) The public updates to a posterior reputation $\rho^+ = \Phi(\rho; a, \iota y)$, where $\iota \in \{0,1\}$ indicates implementation (so the public observation is $(a, \iota y)$).

(vi) Payoffs accrue (defined below).

When strategies are of cutoff type (Definition 3.5), posteriors after each public event do not depend on the realized x beyond the event itself. We denote by

$$\rho_{R,1}(\rho), \quad \rho_{R,0}(\rho), \quad \rho_S(\rho)$$

the Bayes-consistent expected posteriors after, respectively, observing $(a = R, y = 1)$, $(a = R, y = 0)$, and $a = S$, integrating over the truncated signal distributions induced by the strategy and the primitives.

3.3 Payoffs and the success-only bonus

The expert is risk-neutral and derives a reduced-form *career-concerns* payoff from reputation:[6]

$$W : [0,1] \to \mathbb{R}, \qquad W'(\cdot) > 0.$$

[6] $W(\cdot)$ is a reduced-form continuation value (e.g., future wage, promotion probability, or demand for advice) as a function of the posterior reputation. A scale factor κ captures the *strength* of career concerns and drives the corresponding comparative static in Proposition 4.3.

In addition, a *success-contingent bonus* $b \geq 0$ pays only if the expert recommended R, it was implemented, and the outcome succeeded.[7]

Given signal x and reputation ρ, the expert's expected one-period payoff from recommending R versus S is

$$U_R(x;\rho) = \lambda(\rho)\Big[p(x)\,W\Big(\rho_{R,1}(\rho)\Big) + \Big(1-p(x)\Big)\,W\Big(\rho_{R,0}(\rho)\Big)\Big] \;+\; b\,\lambda(\rho)\,p(x), \tag{3.1}$$

$$U_S(x;\rho) = W\Big(\rho_S(\rho)\Big). \tag{3.2}$$

Define the *payoff difference*

$$\Delta(x;\rho) \;\equiv\; U_R(x;\rho) - U_S(x;\rho). \tag{3.3}$$

3.4 Assumptions

Assumption 3.1. *The success probability* $p : \mathbb{R} \to (0,1)$ *is strictly increasing and continuously differentiable in* x *(e.g., induced by an MLRP signal about a binary state).*

Assumption 3.2. *Implementation intensity* $\lambda : [0,1] \to [0,1]$ *is weakly increasing and continuous in reputation* ρ*.*

Assumption 3.3. *The reputational payoff* W *is strictly increasing and continuous in the posterior reputation.*

Assumption RD (Relative diagnosticity) 3.4. For sufficiently high ρ, the expected reputational return to recommending R (relative to S) at the equilibrium cutoff weakly declines in ρ; equivalently, failures are (weakly) more diagnostic than successes when reputation is high and are more likely to be realized when $\lambda(\rho)$ is higher.

For later results, it is convenient to let: (i) a prior parameter $\pi \in (0,1)$ shift success probabilities $p(x;\pi)$ pointwise in x; (ii) an informativeness index θ make $p(\cdot;\theta)$ steeper in the Lehmann sense; (iii) a career-concern strength $\kappa \geq 0$ scale reputation payoffs via $W_\kappa(\cdot) = \kappa W(\cdot)$; and (iv) the bonus $b \geq 0$ enter as in (3.1). Gatekeeping/monitoring will be captured by a parameter T that shifts $\lambda(\rho;T)$ with $\partial\lambda/\partial T \leq 0$.

3.5 Equilibrium

Definition 3.5. A stationary equilibrium consists of a cutoff function $c : [0,1] \to \mathbb{R}$ and a belief-update rule Φ such that:

[7] "Bonus" is shorthand for any success-contingent reward observable to the audience (monetary or otherwise). Examples include public recognition, analyst "star" status, or promotion. Our analysis abstracts from multitasking; in richer environments the instrument can be made small and targeted to risky recommendations.

(a) Given Φ, the expert recommends R if and only if $x \geq c(\rho)$ and S otherwise, for every $\rho \in (0,1)$.

(b) Φ is obtained from Bayes' rule using the cutoff strategy $c(\cdot)$ and the primitives.

Under Assumptions 3.1–3.3, $\Delta(x;\rho)$ in (3.3) is strictly increasing in x, which delivers the cutoff characterization stated formally in Proposition 4.1.

4 Equilibrium and Main Results

The goal of this section is to turn the qualitative forces described in the Introduction into transparent objects that organize the entire analysis. We first show that advice admits a simple global description: a single threshold in the continuous signal separates risky from safe recommendations. This cutoff logic is not a knife–edge artifact; it follows from a familiar single–crossing structure and gives uniqueness and continuity.

We then establish our central comparative static—*reputational conservatism*—and collect additional monotone effects with respect to informativeness, priors, career–concern strength, and the success–only bonus.

Throughout, we keep the presentation intuitive and emphasize how each result maps to observable behavior (frequency of risky recommendations and conditional hit rates).

4.1 Cutoff characterization

Before any design lever is considered, one needs a tractable description of advice. Because higher signals make success more likely, while the safe recommendation carries no signal–contingent reputational risk within the episode, the expert's net gain from risk is strictly increasing in the private signal. This single–crossing property collapses the strategy space to a single number—the cutoff—so that the entire behavior of the expert at a given reputation is summarized by "how much evidence is enough to take risk." This characterization not only yields uniqueness but also anchors all comparative statics that follow.[8]

Proposition 4.1. *Under Assumptions 3.1–3.3, for every reputation level $\rho \in (0,1)$ there exists a (weakly) unique threshold $c(\rho) \in \mathbb{R}$ such that the expert recommends R iff $x \geq c(\rho)$. Moreover,*

$$\partial_x \Delta(x;\rho) = \lambda(\rho)\, p'(x) \Big(W\big(\rho_{R,1}(\rho)\big) - W\big(\rho_{R,0}(\rho)\big) + b \Big) \;>\; 0,$$

so $\Delta(\cdot;\rho)$ is strictly increasing in x.

[8] Uniqueness follows from strict single-crossing of $\Delta(x;\rho)$ in x; multiple equilibria typical of cheap-talk environments do not arise here because actions are verifiable and outcomes feed posteriors directly.

Proof sketch. By (3.1)–(3.3), U_S is independent of x, while U_R increases in x because $p'(x) > 0$ and W is increasing, which implies $W(\rho_{R,1}) > W(\rho_{R,0})$; the bonus adds $b \geq 0$. Hence $\partial_x \Delta > 0$. Since $\Delta(x;\rho) \to -W(\rho_S)$ as $p(x) \to 0$ and $\Delta(x;\rho) \to \lambda(\rho)\, W(\rho_{R,1}) - W(\rho_S) + b\,\lambda(\rho)$ as $p(x) \to 1$, the intermediate value theorem yields a unique root $c(\rho)$. □

4.2 Reputational conservatism and comparative statics

Having pinned down the cutoff, we ask how it moves with reputation and with economically meaningful primitives. Our diagnosticity condition formalizes a common intuition: at higher standing, a visible failure carries more reputational weight than a visible success carries favorable weight.

When implementation is also more likely at higher standing, this makes the downside of a risky call loom larger for respected experts, pushing the cutoff up—hence *reputational conservatism* (Proposition 4.2).[9] Beyond reputation, the model predicts that more informative signals or more favorable priors reduce the evidentiary bar for risk, while stronger career–concern intensity raises it; a success–only bonus shifts the bar down (Proposition 4.3). These effects translate directly into predictions for both the frequency of risky recommendations and their conditional accuracy.

Proposition 4.2. *Suppose Assumption 3.4 holds from some $\bar{\rho} \in (0,1)$ onward. Then the equilibrium cutoff is (weakly) increasing in reputation for $\rho \geq \bar{\rho}$: $c'(\rho) \geq 0$.*

Proof sketch. At interior points, the cutoff satisfies $\Delta(c(\rho);\rho) = 0$. By the implicit function theorem,

$$c'(\rho) \;=\; -\frac{\partial_\rho \Delta(c(\rho);\rho)}{\partial_x \Delta(c(\rho);\rho)}.$$

The denominator is strictly positive by Proposition 4.1. Assumption 3.4 states that at the cutoff the expected reputational return to R relative to S weakly declines with ρ, so $\partial_\rho \Delta(c(\rho);\rho) \leq 0$ for $\rho \geq \bar{\rho}$. Thus $c'(\rho) \geq 0$. □

Proposition 4.3. *Let $p(x;\pi,\theta)$ denote the success probability with prior parameter $\pi \in (0,1)$ and informativeness index θ (higher θ is a Lehmann/Blackwell improvement). Let $W_\kappa = \kappa W$ with $\kappa \geq 0$ scale the career-concerns payoff, and let $b \geq 0$ be the success-only bonus. Then, at interior solutions:*

(i) If θ steepens $p(\cdot;\pi,\theta)$ in the Lehmann sense, then $\partial c/\partial\theta \leq 0$.

(ii) If $p(x;\pi,\theta)$ is pointwise increasing in π, then $\partial c/\partial\pi \leq 0$.

[9] The mechanism is selection: when ρ is high, the public more often sees outcomes of one's risky calls, and—near the decision margin—failures carry more informational weight than successes. The expert internalizes this asymmetric exposure and becomes more conservative.

(iii) Under Assumption 3.4 at moderate/high ρ, $\partial c/\partial \kappa \geq 0$.
(iv) $\partial c/\partial b \leq 0$ for all ρ.

Proof sketch. Differentiate the identity $\Delta(c(\rho;\xi);\rho,\xi) = 0$ w.r.t. each parameter $\xi \in \{\theta, \pi, \kappa, b\}$:

$$\frac{\partial c}{\partial \xi}(\rho;\xi) \;=\; -\frac{\partial_\xi \Delta(c(\rho;\xi);\rho,\xi)}{\partial_x \Delta(c(\rho;\xi);\rho,\xi)}.$$

The denominator is positive by Proposition 4.1. (i)–(ii): U_R is increasing in p while U_S is independent of x; a Lehmann improvement or higher prior raises U_R at any given x, so $\partial_\xi \Delta > 0$ and $\partial c/\partial \xi \leq 0$. (iii): With $W_\kappa = \kappa W$,

$$\partial_\kappa \Delta = \lambda(\rho)\Big[p(c)W(\rho_{R,1}) + \Big(1 - p(c)\Big)W(\rho_{R,0})\Big] - W(\rho_S).$$

Under Assumption 3.4 at the cutoff and for sufficiently high ρ, this is ≤ 0, hence $\partial c/\partial \kappa \geq 0$. (iv): $\partial_b \Delta = \lambda(\rho)\, p(c) > 0$, so $\partial c/\partial b \leq 0$. □

Remark 4.4. The strict monotonicity of $\Delta(\cdot;\rho)$ in x implies the cutoff is unique for each ρ. Under mild regularity on p, λ, W, $c(\rho)$ is continuous in ρ and in the parameters (π, θ, κ, b).

5 Success-Only Bonus and Implementation Intensity

The cutoff logic also makes it straightforward to reason about *policy levers*. We focus on two that are widely used and easy to implement: (i) a success–contingent bonus paid when a risky recommendation is implemented and succeeds, and (ii) gatekeeping/monitoring that affects the probability a risky recommendation is implemented.[10] The first acts directly on the expert's payoff from success; the second alters how often successes and failures become public events. Both levers map cleanly into the cutoff, and thus into experimentation rates, providing practical guidance for committees and principals.

5.1 Bonus → experimentation mapping

A principal often cares about how often experts recommend genuine opportunities rather than about the precise threshold itself. Because the bonus raises the return to a successful risky recommendation uniformly across signals, it shifts the cutoff monotonically and continuously.

This delivers a strictly increasing, one–to–one map from the bonus to the experimentation rate at a given reputation (Proposition 5.1). In practice, this means that a single scalar

[10] Operationally, gatekeeping lowers the arrival rate of observable outcomes. At the margin this raises the evidentiary bar for risk (higher cutoff) and reduces experimentation (Theorem 5.2).

parameter can be tuned to hit a target frequency of risky recommendations without estimating the entire model.

Fix ρ and let F_X denote the (unconditional) CDF of the private signal x. Given the equilibrium cutoff $c(\rho; b)$, define the experimentation rate[11]

$$\varepsilon(\rho; b) \;\equiv\; \mathbb{P}[x \geq c(\rho; b)] \;=\; 1 - F_X\Big(c(\rho; b)\Big).$$

Proposition 5.1. *For any $\rho \in (0, 1)$:*

(i) $\varepsilon(\rho; b)$ is continuous and strictly increasing in $b \geq 0$.

(ii) $\lim_{b\downarrow 0} \varepsilon(\rho; b) =: \varepsilon_0(\rho) \in (0, 1)$ and $\lim_{b\uparrow\infty} \varepsilon(\rho; b) = 1$.

(iii) Hence, for any target $\hat{\varepsilon} \in \Big(\varepsilon_0(\rho), 1\Big)$ there exists a unique $b(\hat{\varepsilon}, \rho)$ implementing $\varepsilon(\rho; b) = \hat{\varepsilon}$.

Proof sketch. By Proposition 4.3(iv), $c(\rho; b)$ is strictly decreasing and continuous in b. Therefore $\varepsilon(\rho; b) = 1 - F_X(c(\rho; b))$ is strictly increasing and continuous. As $b \to \infty$, the b-term in (3.1) dominates, so $c(\rho; b) \downarrow -\infty$ and $\varepsilon(\rho; b) \uparrow 1$. As $b \downarrow 0$, $c(\rho; b) \uparrow c(\rho; 0)$ and the induced rate $\varepsilon_0(\rho) = 1 - F_X(c(\rho; 0))$ lies in $(0, 1)$ under nondegenerate F_X. Strict monotonicity gives uniqueness. □

It is sometimes convenient to invert the cutoff condition at fixed reputation: from $\Delta\Big(c(\rho; b); \rho\Big) = 0$ and (3.3) one obtains the implicit map

$$b\Big(c; \rho\Big) \;=\; \frac{W\Big(\rho_S(\rho)\Big) - \lambda(\rho)\Big[p(c)\, W\Big(\rho_{R,1}(\rho)\Big) + \Big(1 - p(c)\Big)\, W\Big(\rho_{R,0}(\rho)\Big)\Big]}{\lambda(\rho)\, p(c)}. \tag{5.1}$$

Combining $c = F_X^{-1}(1 - \hat{\varepsilon})$ with (5.1) yields the unique $b(\hat{\varepsilon}, \rho)$.

5.2 Gatekeeping, monitoring, and implementation intensity

Gatekeeping and monitoring change how often risky recommendations are actually put into action and thus how frequently outcomes become public signals about ability.

When implementation becomes stricter, both successes and failures are observed less often, but at the cutoff the net effect is to raise the evidentiary bar for taking risk. Consequently, stricter gatekeeping reduces experimentation in a predictable way (Corollary 5.2). This provides a simple organizational insight: approval processes that throttle implementation will mechanically select toward safer advice, even holding information quality fixed.

[11] $\varepsilon(\rho; b)$ is the ex-ante probability of a risky recommendation at reputation ρ under bonus b. It is not a hazard rate; its comparative statics follow directly from the cutoff's monotonicity in Theorem 4.3 and the mapping in Theorem 5.1.

Let T index gatekeeping/monitoring stringency so that $\partial\lambda(\rho;T)/\partial T \le 0$ (higher T means stricter implementation). Differentiate $\Delta\Big(c(\rho;b);\rho,\lambda(\rho;T)\Big) = 0$ to obtain

$$\frac{\partial c}{\partial\lambda}(\rho;b) \;=\; -\frac{\partial_\lambda\Delta}{\partial_x\Delta} \;=\; -\frac{\Big[p(c)\,W(\rho_{R,1}) + \Big(1-p(c)\Big)\,W(\rho_{R,0})\Big] + b\,p(c)}{\lambda(\rho)\,p'(c)\,\Big(W(\rho_{R,1}) - W(\rho_{R,0}) + b\Big)} \tag{5.2}$$

where all objects are evaluated at $x = c(\rho;b)$ and we suppress the ρ-arguments for brevity.

Corollary 5.2. *If the bracketed term in the numerator of* (5.2) *is nonnegative (in particular, for any $b > 0$), then $\partial c/\partial\lambda \le 0$. Consequently, with $\partial\lambda/\partial T \le 0$,*

$$\frac{\partial c}{\partial T} \;\ge\; 0 \qquad \text{and} \qquad \frac{\partial\varepsilon}{\partial T} \;=\; -f_X(c)\,\frac{\partial c}{\partial T} \;\le\; 0,$$

so stricter gatekeeping (higher T) raises the cutoff and lowers experimentation.

Proof sketch. The denominator of (5.2) is positive by Proposition 4.1. If the numerator is nonnegative (always true when $b > 0$), then $\partial c/\partial\lambda \le 0$; the sign for T follows by the chain rule and $\varepsilon = 1 - F_X(c)$. □

6 Discussion and Extensions

The model clarifies a common pattern in advisory settings: elites look cautious not (only) because they are inherently risk-averse, but because visible failures are more informative and more likely to be realized when one's recommendations are routinely implemented.

Two organizational levers follow immediately. First, success–contingent bonuses offer a *direct* way to tune experimentation without wholesale contract redesign; the one–to–one bonus–to–frequency map in Proposition 5.1 makes the tool easy to calibrate. Second, gatekeeping and approval processes should be viewed as *selection devices*: tightening them shifts the institution toward fewer risky recommendations, which can be beneficial when failures are very costly but harmful when exploration is valuable.

In committees, these levers can be combined—e.g., relax approval for high–priority domains while raising success bonuses, and do the opposite for low–priority domains.[12]

Three predictions travel well across domains.

(i) *Selection on risk by standing:* holding information quality fixed, higher–reputation experts issue fewer risky recommendations but enjoy higher conditional hit rates.

[12] A practical rule of thumb from Theorem 5.1: fix a target experimentation frequency by domain and reputation tier, compute $b(\hat{\varepsilon},\rho)$ via (5.1), and recalibrate periodically as diagnostics (signal informativeness) evolve.

(ii) *Information quality effects:* when signals become more informative (better diagnostics, more granular evidence), cutoffs fall and both experimentation and conditional accuracy can move in the same direction.

(iii) *Policy invariants:* raising the success–only bonus increases experimentation at every reputation level; stricter gatekeeping reduces it.

These predictions are qualitative and do not rely on parametric structure beyond single–crossing and RD.

Whether conservatism is beneficial depends on primitives outside our informational core: the social value of success, the cost of failure, and how often implementation should occur. If failures impose large external costs, conservatism induced by career concerns can be partially corrective; if successes generate spillovers or learning, it can be inefficiently stifling.

We intentionally keep welfare analysis light, as it quickly becomes environment-specific; nonetheless, Proposition 5.1 provides a tractable instrument to align individual incentives with institutional objectives.

Our results rest on three elements: continuous private information satisfying single–crossing (MLRP), a monotone implementation probability in reputation, and a reputational payoff increasing in posterior beliefs. The cutoff characterization and the comparative statics survive standard perturbations (alternative continuous families, smooth changes in W or λ). Assumption RD is microfounded in Gaussian settings (Appendix B) and is consistent with the idea that failures near the decision margin are especially revealing.

Promising directions include: committees with multiple advisors (thresholds become profile-dependent and may generate strategic complementarities), multidimensional signals (cutoffs become surfaces with potentially interesting geometry), and endogenous implementation (where committees choose λ jointly with incentives). Each preserves the core selection logic while introducing new levers—e.g., rotating gatekeeping responsibilities or state–contingent bonuses. We leave these to future work.

7 Conclusion

We studied expert recommendations under career concerns in a continuous–signal environment with observable outcomes. Three takeaways emerge. First, *tractability*: advice admits a global cutoff characterization that is unique and continuous, providing a simple summary statistic—“how much evidence is enough for risk?”—for any given reputation level. Second, *reputational conservatism*: under a mild diagnosticity condition, the cutoff (weakly) increases with reputation, implying fewer risky recommendations but higher conditional hit rates among high–reputation experts. Third, *implementable levers*: a success–only bonus generates a strictly increasing, continuous map to experimentation, while gatekeeping shifts selection on risk via implementation intensity.

These insights are portable and testable. They rationalize widely noted patterns in advisory organizations and furnish principals with a minimal tool to steer behavior without micromanaging information production or rewriting complex contracts. The theory deliberately focuses on a one–episode benchmark; extending the cutoff logic to committees, repeated interactions, and endogenous implementation is a natural agenda.

Our hope is that the combination of a clean behavioral summary (the cutoff), clear qualitative predictions, and concrete levers will be useful to both theorists studying reputational incentives and practitioners designing expert systems.

Data and Code Availability

This research is purely theoretical and does not use data, code, or experimental materials. No replication files are necessary.

References

Azrieli, Yaron (2021). “Monitoring Experts”. In: *Theoretical Economics* 16, pp. 1313–1350.

Catonini, Emiliano and Sergey Stepanov (2023). “Reputation and Information Aggregation”. In: *Journal of Economic Behavior & Organization* 208, pp. 156–173.

Dworczak, Piotr and Alessandro Pavan (2022). “Preparing for the Worst but Hoping for the Best: Robust (Bayesian) Persuasion”. In: *Econometrica* 90.5, pp. 2017–2051.

Escobar, Juan F. and Qiaoxi Zhang (2021). “Delegating Learning”. In: *Theoretical Economics* 16.2, pp. 571–603.

Halac, Marina and Ilan Kremer (2020). “Experimenting with Career Concerns”. In: *American Economic Journal: Microeconomics* 12.1, pp. 260–288.

Kamenica, Emir and Matthew Gentzkow (2011). “Bayesian Persuasion”. In: *American Economic Review* 101.6, pp. 2590–2615.

Klein, Nicolas and Tymofiy Mylovanov (2017). “Will Truth Out?—An Advisor’s Quest to Appear Competent”. In: *Journal of Mathematical Economics* 72, pp. 112–121.

Kolotilin, Anton (2015). “Experimental Design to Persuade”. In: *Games and Economic Behavior* 90, pp. 215–226.

Kolotilin, Anton, Tymofiy Mylovanov, Andriy Zapechelnyuk, and Ming Li (2017). “Persuasion of a Privately Informed Receiver”. In: *Econometrica* 85.6, pp. 1949–1964.

Liu, Yaozhou Franklin and Amal Sanyal (2012). “When Second Opinions Hurt: A Model of Expert Advice under Career Concerns”. In: *Journal of Economic Behavior & Organization* 84.1, pp. 1–16.

Marinovic, Iván and Martin Szydlowski (2022). “Monitoring with Career Concerns”. In: *The RAND Journal of Economics* 53.2, pp. 404–428.

Moroni, Sofia (2022). “Experimentation in Organizations”. In: *Theoretical Economics* 17.3, pp. 1403–1450.

Okat, Deniz and John G. F. Nash (2024). “Delegating Trial and Error”. In: *Journal of Economic Theory* 217, p. 105802.

Ottaviani, Marco and Peter Sørensen (2006). “Reputational Cheap Talk”. In: *The RAND Journal of Economics* 37.1, pp. 194–210.

Ottaviani, Marco and Peter Norman Sørensen (2006). “Professional Advice”. In: *Journal of Economic Theory* 126.1, pp. 120–142.

Wittbrodt, Cole and Nathan Yoder (2025). “Delegating Experiments”. SSRN Working Paper No. 5221631, April 17, 2025.

A Proofs

Lemma A.1. *Under Assumptions 3.1–3.3, for any fixed* $\rho \in (0,1)$ *the payoff difference* $\Delta(x;\rho) = U_R(x;\rho) - U_S(x;\rho)$ *is strictly increasing in* x*. Moreover,*

$$\partial_x \Delta(x;\rho) = \lambda(\rho)\, p'(x) \Big(W(\rho_{R,1}(\rho)) - W(\rho_{R,0}(\rho)) + b \Big) \;>\; 0.$$

Proof. From (3.1)–(3.2), U_S is independent of x and $\partial_x U_R = \lambda(\rho)\, p'(x) \Big(W(\rho_{R,1}) - W(\rho_{R,0}) + b \Big)$. By Assumption 3.1, $p'(x) > 0$; by Assumption 3.3, W is strictly increasing so $W(\rho_{R,1}) > W(\rho_{R,0})$; and $b \geq 0$. Thus $\partial_x \Delta > 0$. □

Proof of Proposition 4.1. Fix ρ. By Lemma A.1, $\Delta(\cdot;\rho)$ is strictly increasing, hence has at most one root. For x with $p(x) \approx 0$, $\Delta(x;\rho) \approx -W(\rho_S) < 0$. For x with $p(x) \approx 1$, $\Delta(x;\rho) \approx \lambda(\rho) W(\rho_{R,1}) - W(\rho_S) + b\,\lambda(\rho)$. Since $W(\rho_{R,1}) \geq W(\rho_S)$ and $b \geq 0$, the latter is ≥ 0. By continuity, there exists a (weakly) unique $c(\rho)$ with $\Delta(c(\rho);\rho) = 0$. □

Proof of Proposition 4.2. At interior points, the cutoff satisfies $\Delta(c(\rho);\rho) = 0$. By the implicit function theorem,

$$c'(\rho) \;=\; -\frac{\partial_\rho \Delta(c(\rho);\rho)}{\partial_x \Delta(c(\rho);\rho)}.$$

The denominator is strictly positive by Lemma A.1. Assumption 3.4 states that at the cutoff the expected reputational return to R (relative to S) weakly declines with ρ for $\rho \geq \bar{\rho}$, i.e. $\partial_\rho \Delta(c(\rho);\rho) \leq 0$. Hence $c'(\rho) \geq 0$ on $[\bar{\rho}, 1)$. □

Proof of Proposition 4.3. Let $\xi \in \{\theta, \pi, \kappa, b\}$ and write the indifference condition $\Delta(c(\rho;\xi);\rho,\xi) = 0$. Differentiate:

$$\frac{\partial c}{\partial \xi}(\rho;\xi) \;=\; -\frac{\partial_\xi \Delta(c(\rho;\xi);\rho,\xi)}{\partial_x \Delta(c(\rho;\xi);\rho,\xi)}.$$

The denominator is > 0 by Lemma A.1. (i) Lehmann improvements steepen $p(\cdot;\pi,\theta)$, so U_R rises pointwise in x, implying $\partial_\theta \Delta > 0$ and $\partial c/\partial\theta \leq 0$. (ii) If $p(\cdot;\pi,\theta)$ increases pointwise in π, then $\partial_\pi \Delta > 0$ and $\partial c/\partial \pi \leq 0$. (iii) With $W_\kappa = \kappa W$, $\partial_\kappa \Delta = \lambda(\rho)\Big[p(c) W(\rho_{R,1}) + (1 - p(c)) W(\rho_{R,0})\Big] - W(\rho_S)$. Under Assumption 3.4 at the cutoff and for moderate/high ρ, this is ≤ 0, hence $\partial c/\partial\kappa \geq 0$. (iv) $\partial_b \Delta = \lambda(\rho)\, p(c) > 0$, so $\partial c/\partial b \leq 0$. □

Proof of Proposition 5.1. By Proposition 4.3(iv), $c(\rho;b)$ is strictly decreasing and continuous in b. Thus $\varepsilon(\rho;b) = 1 - F_X(c(\rho;b))$ is strictly increasing and continuous in b. As $b \downarrow 0$, $c(\rho;b) \uparrow c(\rho;0)$ so $\varepsilon(\rho;b) \to \varepsilon_0(\rho) := 1 - F_X(c(\rho;0)) \in (0,1)$ under nondegenerate F_X. As $b \uparrow \infty$, the b-term in (3.1) dominates, forcing $c(\rho;b) \downarrow -\infty$ and $\varepsilon(\rho;b) \uparrow 1$. Strict monotonicity yields the uniqueness of $b(\hat{\varepsilon}, \rho)$ for any $\hat{\varepsilon} \in (\varepsilon_0(\rho), 1)$. □

Proof of Corollary 5.2. Differentiate $\Delta(c(\rho; b); \rho, \lambda) = 0$ w.r.t. λ to obtain (5.2). The denominator is positive by Lemma A.1. If the numerator is nonnegative (e.g., whenever $b > 0$), then $\partial c/\partial\lambda \leq 0$. Since $\partial\lambda/\partial T \leq 0$, the chain rule gives $\partial c/\partial T \geq 0$. Finally, $\varepsilon = 1 - F_X(c)$ implies $\partial\varepsilon/\partial T = -f_X(c)\,\partial c/\partial T \leq 0$. □

Continuity note. Under the standing assumptions and mild regularity of F_X, standard arguments imply that $c(\rho)$ is continuous in ρ and in (π, θ, κ, b); details are omitted. □

B Gaussian Microfoundation for Assumption RD

This appendix sketches conditions under which Assumption 3.4 holds in a canonical Gaussian environment.

Setup. Let the payoff state be $s \in \{0, 1\}$ with prior $\pi \in (0, 1)$. The expert privately observes a signal $x \in \mathbb{R}$ with

$$x \mid s \sim \mathcal{N}(\mu_s, \sigma^2), \qquad \mu_1 > \mu_0,$$

so that $p(x) = \mathbb{P}[s = 1 \mid x]$ is strictly increasing in x (MLRP). Suppose that the market holds public reputation ρ about the expert (e.g., the probability of high ability), and $\lambda = \lambda(\rho)$ is weakly increasing in ρ. Given a cutoff strategy $a = R$ iff $x \geq c(\rho)$, the public posteriors after (a, y) are $\rho_{R,1}(\rho)$, $\rho_{R,0}(\rho)$, and $\rho_S(\rho)$, obtained by Bayes' rule integrating over the truncated normal $x \mid x \geq c(\rho)$.

Claim. For sufficiently high ρ, the expected reputational return to a risky recommendation (relative to a safe one) at the cutoff weakly declines in ρ:

$$\frac{\partial}{\partial\rho}\Big\{\lambda(\rho)\Big[p(c(\rho))\,W(\rho_{R,1}(\rho)) + (1 - p(c(\rho)))\,W(\rho_{R,0}(\rho))\Big] \;-\; W(\rho_S(\rho))\Big\} \;\leq\; 0.$$

Sketch of argument. Two forces drive the derivative:

(i) *Selection (diagnosticity) effect.* Conditional on $a = R$, the signal is truncated to $x \geq c(\rho)$. At higher ρ, equilibrium behavior (by monotonicity of best responses in Gaussian MLRP settings) implies a (weakly) higher cutoff $c(\rho)$ in the relevant region; the truncated distribution then places more mass on high x, making *failure* $y = 0$ at the cutoff increasingly unlikely ex ante. Therefore, the likelihood ratio of observing $y = 0$ vs. $y = 1$ after $a = R$ grows (locally) with ρ, and, because W is increasing, the loss from a failure relative to the gain from a success becomes larger in expectation.

(ii) *Implementation effect.* Since $\lambda'(\rho) \geq 0$, risky recommendations are implemented more often when ρ is high, so both successes and failures realize more frequently. At the cutoff, the success probability is $p(c(\rho)) \in (0, 1)$. With even a small bonus $b \geq 0$, the reputational-plus-bonus value of a success minus a failure is positive, but the *expected* reputational component at the cutoff is dominated by the failure term because failures are rare but highly diagnostic under truncation.

Formally, differentiating the expected reputational component and using properties of truncated normals (monotone hazard) yields a nonpositive derivative for sufficiently high ρ, establishing Assumption 3.4. This microfoundation is standard in Gaussian single-crossing environments; full algebra is omitted for brevity. □