No. 2026-012

Is Solving a Proxy for Understanding?
A Performance-Measurement Analysis of
Goal Misalignment in AI-Generated Mathematics

October 19, 2026

Takanori Suzuki

Is Solving a Proxy for Understanding? A Performance-Measurement Analysis of Goal Misalignment in AI-Generated Mathematics

Takanori Suzuki*

October 19, 2026

Research Institute of Business Administration (Sanken), Waseda University
Working Paper Series, No. 2026-012

Working paper note. English version of the author's Japanese-language manuscript with the same title, prepared for *The Waseda Commercial Review* (*Waseda Shogaku*), No. 473 (December 2026). The models, results, and proofs are shared with that manuscript. Information cutoff: October 18, 2026.

Abstract

This paper examines goal misalignment between AI labs and the mathematical community through performance-measurement theory. Motivated by a 2026 declaration emphasizing conceptual understanding, Model I distinguishes effort toward solving from support for human understanding. When the tasks are substitutes, market rewards for solving crowd out such support. Norms can improve social surplus, but measurement noise, apparent compliance, and the option to publish outside the community channel constrain their effectiveness. The main analysis shows that, under the stated conditions, labs with stronger solving incentives are more likely to exit and face weaker effective norms despite a greater need for corrective incentives. Partial compliance and a common norm for heterogeneous labs introduce further trade-offs. Model II separately studies incremental learning from current human corrections. A positive marginal cost at zero correction leaves a residual gap; additional learning channels can eliminate it under a cumulative-learning condition. If correction is the sole means of maintaining human mathematical taste, skill depreciation can generate lock-in. Benefits to future users also imply socially insufficient correction. These conditional results inform an interpretation of emerging release norms as combinations of performance measures, disclosure, and certification. Implications concern research institutions, education, and management accounting; the models do not establish general limits on AI capabilities.

Keywords: performance measurement; proxy measures; multitask incentives; AI-generated mathematics; formal verification; verifiability; understandability; participation constraints; correction-based learning; release norms
JEL classification: D82, D83, M41, O33.

*Faculty of Commerce, Waseda University. Email: suzuki@waseda.jp. ORCID: 0009-0008-8955-7003.

1 Introduction

1.1 The problem

As AI reduces the cost of generating mathematical results, the costs of verifying and understanding new proofs and manuscripts need not fall at the same rate. This asymmetry between generation and verification has been discussed as a problem for science more generally, not only for mathematics (Ma, 2026).

Since September 2026, this asymmetry has manifested itself in specific events in mathematics (Section 2). On September 8, OpenAI claimed to have produced an example of the three-dimensional incompressible Navier–Stokes equations with forcing for which a smooth solution does not exist globally in time (OpenAI, 2026a). The claim is accompanied by a Lean formalization, but it has not undergone peer review, and questions have been raised about whether the Lean proof corresponds to the natural-language proof (Bastounis et al., 2026). On September 11, a joint declaration entitled "A Severe Misalignment of AI in Mathematics" was published, with twenty-five Fields Medalists as its initial signatories (Avila et al., 2026). The declaration argues that AI companies' pursuit of mathematical problems as benchmarks is harming mathematics as a discipline, stating:

> "solving problems is only a tool and proxy for achieving the primary goal of conceptual understanding and insight."

On September 29, the Advisory Group on Mathematics and Artificial Intelligence (AGMAI) issued guidelines on the responsible release of AI-generated mathematics (AGMAI, 2026a). On October 6, however, OpenAI released more than 700 manuscripts produced by an unreleased internal model on GitHub, without going through journals (OpenAI, 2026c,d). AGMAI issued a statement that day (AGMAI, 2026b). The following day, the researchers' organization AHM criticized the release and called for a halt to cooperation with OpenAI (AHM, 2026).

The declaration's use of *misalignment* does not refer to the problem of aligning AI behavior with human intentions. It refers to a divergence between the goals of AI companies and those of the mathematical community. AI companies compete over "problems solved," which are rewarded by the market—investors, users, and the media. From the declaration's perspective, by contrast, the mathematical community's primary goal is conceptual understanding. This paper takes that to be the community's objective; widespread endorsement does not imply that every member of the community shares the same objective. The possibility that rewarding an easily measured indicator undermines a difficult-to-measure objective is a classic concern of performance-measurement theory (Kerr, 1975; Holmström and Milgrom, 1991). This paper formalizes that structure within the framework of performance measurement and evaluates the norms for releasing results that emerged in 2026.

1.2 Research questions

The paper addresses three research questions.

RQ1 How can the divergence between the goals of AI companies and those of the mathematical community be formulated as a problem of performance measurement?

RQ2 Under what conditions does an AI that learns from human corrections converge to human evaluative judgment, or mathematical taste, and where does a gap remain?

RQ3 To what extent can the norms for releasing results that emerged in 2026 correct these divergences?

RQ1 takes the declaration's claim itself as the object of analysis. RQ2 asks how far an AI that supports human understanding can learn human judgment. The quality of such support is difficult to verify, and the AI learns it through human corrections. RQ3 uses the analyses of RQ1 and RQ2 to evaluate actual institutional responses.

1.3 Contributions

The paper makes three contributions.

(1) It formulates the declaration's claim as a multitask problem in which incentives on measured tasks crowd out unmeasured tasks (Section 4). Propositions 4.1 and 4.2 are applications of existing multitask models. The central theoretical contribution is the analysis of the enforcement capacity of norms when labs can release results outside them (Propositions 4.3–4.5). The analysis identifies a selection pattern in which the labs most in need of correction are those that exit the norm and face only a weak effective norm. The paper also draws on the qualitative characteristics in accounting's conceptual framework to distinguish the verifiability provided by formal verification from faithful representation and understandability (Section 3).

(2) For an AI whose learning is tied to current human corrections, the paper establishes propositions on a lower bound on the gap, necessary and sufficient conditions for the gap to converge to zero, lock-in caused by the depreciation of human taste, and underprovision of corrections due to learning externalities (Section 5).

(3) Taking institutional developments in 2026 as a case, the paper maps individual provisions of norms for releasing results to model parameters and discusses their corrective effects and limitations (Section 6).

The analysis in Propositions 4.3–4.5 is related to research on certification and self-regulation. Lizzeri (1999) analyzes how much information a certification intermediary discloses, while Lerner and Tirole (2006) analyzes providers' choice of certifier for their ideas or products (forum shopping). DeMarzo et al. (2005) shows that self-regulatory organizations relax enforcement in their members' interests and that the threat of government enforcement tightens it. The labs in this paper also face a related choice of whether to participate in or exit a certification system, because they can choose between a community channel subject to the norm and their own channel. The analysis has two distinctive features. First, the task measured by the norm is support for human understanding, which is difficult to verify, and the option to exit interacts with multitask crowding out. As long as a lab remains subject to the norm, the norm's effect diminishes as the measure becomes harder to verify (Proposition 4.2). When exit is possible, however, cheaper apparent compliance expands the set of participants, and this relationship can be reversed (Appendix 4.B.1). Second, the paper considers both partial exposure to evaluation outside the norm and a common norm for labs facing different intensities of competition. It shows how the selection whereby the labs most in need of correction exit the norm affects the design of the norm. A rigorous comparison with these earlier studies remains a task for future research.

1.4 Organization of the paper

Section 2 sets out the sequence of events in 2026 and the nature of the sources. Section 3 presents the conceptual framework. Sections 4 and 5 analyze Model I and Model II, respectively, and Section 6 evaluates the norms for releasing results. The two models are not integrated into a single model. They separately analyze two mechanisms that impede support for human understanding: crowding out by market incentives and limits to learning from corrections. Section 7 discusses implications for commerce and accounting, and Section 8 concludes and sets out limitations. Proofs of the propositions and details of the numerical examples are provided in the appendices. The factual account relies on material publicly available as of the information cutoff specified in Section 2. Ongoing developments, particularly claims still undergoing verification, are described as they stood at that date.

2 Events and sources

This section sets out the main events of 2026 and the nature of the sources used in this paper, providing the factual basis for the subsequent analysis. The account draws on material publicly available as of the information cutoff, October 18, 2026.

2.1 Main events

Table 1 presents the main events of 2026.

Table 1: Main events in 2026

Date	Event	Source
June 2	The Leiden Declaration on Artificial Intelligence and Mathematics addresses disclosure of AI use, human responsibility for results, attribution, the need for public alternatives to proprietary technologies, and autonomy over the direction of research.	Leiden Declaration (2026)
September 8	OpenAI claims to have produced an example of the three-dimensional incompressible Navier–Stokes equations with forcing for which a smooth solution does not exist globally in time, accompanied by a Lean formalization.	OpenAI (2026a)
September 11	A joint declaration is published with twenty-five Fields Medalists as its initial signatories. The Clay Mathematics Institute issues a statement the same day indicating that it will allow the verification process to proceed without haste.	Avila et al. (2026); Clay Mathematics Institute (2026)
September 21	OpenAI announces the formation of an advisory group (AGMAI), based at the Institute for Advanced Study (IAS) and operating independently of OpenAI.	OpenAI (2026b)
September 29	AGMAI publishes its guidelines, "Responsible Release of AI-Generated Mathematics."	AGMAI (2026a)

(Table 1 continued)

Date	Event	Source
October 1	Citing, among other reasons, a surge in low-value submissions made using AI, arXiv revises its submission limits to apply to all submitters and all subject categories.	Boboris (2026)
October 6	OpenAI releases manuscripts produced by an unreleased internal model on GitHub. AGMAI issues a statement the same day. A paper on the Navier–Stokes claim is published arguing that the Lean proof does not correspond to the natural-language proof.	OpenAI (2026c,d); AGMAI (2026b); Bastounis et al. (2026)
October 7	OpenAI withdraws three manuscripts and revises fourteen. The researchers' organization AHM issues a statement criticizing the release and calling for a halt to cooperation with OpenAI.	OpenAI (2026d); AHM (2026)
October 15	The Simons Foundation announces a $15 million initiative in combined funding and resources, including contributions from Anthropic and OpenAI. The foundation will allocate the funding, including through an open call.	Simons Foundation (2026)
October 16	Seven authors, including Avigad and Buzzard, discuss trade-offs among speed, scale, and understanding in response to proposals to automatically formalize the mathematical literature as a whole.	Avigad et al. (2026)

2.2 The joint declaration and the guidelines

The joint declaration of September 11 was published on a dedicated website, with twenty-five Fields Medalists, including T. Tao, as its initial signatories (Avila et al., 2026). As of the information cutoff, the website listed twenty-eight Fields Medalists and more than 8,300 endorsers. The declaration addresses premature announcements without adequate write-ups, attribution and plagiarism, students and the human transmission of knowledge, and the importance of understanding. It does not address formal verification. Nor does it call for a ban on AI use: it acknowledges AI's potential to enhance and accelerate genuine mathematical research and understanding. It nevertheless calls on the mathematical community, AI companies, and society to address these issues urgently. The Japanese translation of the declaration renders the sentence quoted in Section 1 in terms equivalent to "solving problems is only a tool and an indirect indicator for achieving the primary goal of conceptual understanding and insight."

AGMAI consists of nine mathematicians who came together after OpenAI approached some of its members about establishing an external advisory body. It states that it operates independently of AI companies and without compensation (AGMAI, 2026c; OpenAI, 2026b). The September 29 guidelines (AGMAI, 2026a) state that the release of results understood by humans is subject to normal practices, such as peer review in journals. They then set out the following requirements, mainly for the release of results that no human yet understands. The final item below comes from a separate section on

broad access to models.

- The guidelines do not endorse some AI labs' practice of testing proprietary, unreleased models on advanced mathematical problems and call for this practice to stop.
- AI labs that release large quantities of mathematical results without human understanding bear responsibility for ensuring that human understanding follows. The development of that understanding should, however, be led by the community, not by AI labs, including the lab that produced the results.
- Releases should include a search of the prior literature and appropriate attribution; a readable paper format, with an introduction and precise statements of theorems and proofs; deposit in an academic repository not controlled by an AI lab; avoidance of using the release to advertise a model; disclosure of the model name, prompts, summaries of reasoning, time taken, and computational cost; and formalization wherever possible. When many results are released at once, a document should report the number of problems of comparable difficulty that were not solved and explain how the problems were selected.
- Labs should support workshops and summer schools for human understanding, as well as the writing of books and extended expository articles by experts. Decisions on allocating the funding should be entrusted to established nonprofit organizations that already have funding-allocation procedures.
- The mathematical community should be assured broad and equitable access to publicly available models, avoiding a two-tier structure in which labs pull ahead of other researchers.

2.3 The October 6 release and subsequent developments

On October 6, OpenAI released a collection of manuscripts produced by an unreleased internal model in a GitHub repository, without going through journals (OpenAI, 2026c,d). The initial release comprised 722 manuscripts across 372 families. On October 7, three were withdrawn because of sign errors and other issues, and fourteen were revised, leaving 719 manuscripts as of the information cutoff. The main results of approximately 42% of the manuscripts (300/719) were formalized in Lean, and summaries of reasoning were released for ten families. The model was reportedly given approximately 4,000 problems. OpenAI states that it continues to explore community-led release channels consistent with the guidelines. The release conflicts with the guidelines' call to stop testing proprietary models on advanced problems. At the same time, some of the requested disclosures were provided, including formalizations, summaries of reasoning, estimates of computational costs, and the number of problems attempted.

In its statement that day, AGMAI described the release as the beginning, rather than the completion, of the process of human understanding, and reaffirmed its guidelines (AGMAI, 2026b). It also stated that its advisory role should not be interpreted as a judgment on the significance of the results or as an endorsement of the process through which OpenAI obtained them; ultimately, assessment of the extent of compliance with the guidelines was left to the mathematical community. In its statement the following day, the researchers' organization AHM criticized the release of more than 700 files at once as a demonstration of power rather than an act of scholarship. It stated that OpenAI had disregarded a central premise of the advisory group's position and called

on mathematicians to cease cooperation with OpenAI (AHM, 2026).

In an essay published on October 16, seven researchers in formalization argued, in response to proposals to automatically formalize the entire mathematical literature, that poor-quality formalization is worse than none (Avigad et al., 2026). The same essay noted that the natural-language and formalized versions of many of the released manuscripts diverged and that many of the natural-language write-ups were impenetrable. A paper on the Navier–Stokes claim argued that the Lean proof did not correspond to the natural-language proof (Bastounis et al., 2026). That paper does not make a claim about the correctness of the natural-language proof itself. The seven authors of the essay take the view that there is now little doubt about the correctness of the result, even though the natural-language proof is mostly incomprehensible (Avigad et al., 2026). The Clay Mathematics Institute has stated that it will allow the verification process to proceed without haste (Clay Mathematics Institute, 2026).

Responses from the community have also begun. Citing, among other reasons, the ability to submit large numbers of low-value papers using AI, arXiv revised its submission limits, which had previously depended mainly on moderators' discretion, to apply to all submitters and all subject categories, with a limit of two new submissions per month (Boboris, 2026). The Simons Foundation launched an initiative including contributions from Anthropic and OpenAI and stated that the foundation would allocate the funding, including through an open call (Simons Foundation, 2026).

2.4 Nature of the sources and information cutoff

The sources used in this paper are of several different kinds. The first consists of declarations with identified signatories: the joint declaration and the Leiden Declaration. The second consists of guidelines and statements by an advisory group (AGMAI), a statement by a researchers' organization (AHM), and a statement by a research institution (the Clay Mathematics Institute). The third consists of documents and a repository published by a company itself (OpenAI). The fourth consists of researchers' essays and papers, including contributions to blogs and papers on arXiv. The paper relies on primary sources and seeks to avoid reliance on news reports. It reads the wording of declarations, guidelines, and statements as normative documents, much as in a comparative analysis of accounting standards.

Because developments are occurring on a daily basis, the factual account is confined to the information cutoff of October 18, 2026. In particular, the Navier–Stokes claim is treated as a claim under continuing verification, not as an established result. Where figures vary over time, as with the number of manuscripts, the paper uses the figures in primary sources as of the cutoff and notes how they changed.

3 Conceptual framework

This section clarifies the concepts used in the analysis from Section 4 onward. First, it interprets the declaration's statement that solving is only a proxy for understanding as a proxy-measure problem in performance-measurement theory (Section 3.1). Second, it reviews the discussion of the relationship between proof and understanding in mathematics (Section 3.2). Third, it uses the qualitative characteristics in accounting's conceptual framework to distinguish what formal verification does and does not provide

(Section 3.3). Fourth, it explains the need to distinguish global from local understanding (Section 3.4).

3.1 Solving as a proxy measure

When organizations cannot measure their objective directly, they reward a measure correlated with it. Rewarded behavior may then diverge from desired behavior. Kerr (1975) calls this "the folly of rewarding A, while hoping for B." Strathern (1997)'s formulation of what is known as Goodhart's law—that a measure ceases to function as a measure when it becomes a target—addresses the same phenomenon.

Performance-measurement theory has made this problem more precise in the language of contract theory. In the multitask model of Holmström and Milgrom (1991), an agent allocates effort across several tasks, only some of which are measured. If the tasks are substitutes in cost, strengthening incentives on the measured task draws effort away from unmeasured tasks. The more important the unmeasured tasks are, therefore, the weaker the incentives on the measured task should be. Job design that assigns difficult-to-measure tasks to a separate party can also mitigate incentive distortions. Feltham and Xie (1994) defines the extent to which a measure points in the same direction as the objective as *congruity* and identifies it, along with noise in the measure, as a determinant of the measure's value. Baker (1992) shows that the efficiency of incentive contracts depends on the correlation between the effects of effort on the measure and on the objective. Baker (2002) calls the divergence between the measure and the objective *distortion* and establishes a trade-off between distortion and risk. Datar et al. (2001) shows that optimal weights on multiple measures balance the congruity of the composite measure as a whole with the objective against the risk that the measures impose on the agent.

In the language of performance measurement, problems solved as benchmarks are a measure of AI's mathematical capability, and the market—investors, users, and the media—rewards that measure. The declaration's criticism (Avila et al., 2026) can be read as directed at these incentives. Following the declaration's position, this paper takes conceptual understanding to be the mathematical community's objective. The declaration's statement that solving is only a proxy for understanding can be read as identifying low congruity between the measure and the objective. Section 4 formalizes this interpretation in a multitask model. In this paper, *crowding out* refers to the substitution of effort across tasks. It is distinct from the crowding out of intrinsic motivation by extrinsic rewards (Frey and Jegen, 2001).

3.2 Proof and understanding

The relationship between proof and understanding in mathematics was discussed well before the advent of AI. Thurston (1994) argues that mathematicians produce human understanding, rather than a quantity of theorems and proofs, and that success is measured by whether their work enables people to think more clearly about mathematics. De Millo et al. (1979) argues that acceptance of a proof as correct arises through social processes in which proofs are read, discussed, and used, rather than primarily through formal verification.

The search for proofs by AI belongs to what Reichenbach (1938) calls the context of discovery. Verification by formal proof assistants such as Lean addresses one part of

the context of justification: the correctness of derivation. Understanding, however, is distinct from both. Even when a proof has been found and its correctness established, humans may not grasp why it works.

This distinction became a practical issue in 2026. Even when an AI-generated proof is formalized in Lean, type checking guarantees only the correctness of the derivation in terms of the formalized statements. Whether those statements faithfully represent the intended problem, and whether the formal proof corresponds to the natural-language proof, require separate checks. In the case of OpenAI's Navier–Stokes claim, it has been argued that the Lean proof does not correspond to the natural-language proof (Bastounis et al., 2026). Research on automatically generating human-oriented structures (blueprints) from Lean formalizations reports a case in which the proof of an AI-generated lemma did not depend on the surrounding lemmas, revealing that the statement itself was semantically incorrect (Zhu et al., 2026). Research has also emerged on rewriting formal proofs according to measures such as readability (Ahuja et al., 2025), alongside discussion of trade-offs among speed, scale, and understanding in proposals to automatically formalize the mathematical literature as a whole (Avigad et al., 2026). As the cost of generation falls sharply, how to allocate the costs of verification and understanding has become a problem for science more generally, not only for mathematics (Ma, 2026).

3.3 Verifiability, faithful representation, and understandability

Accounting's conceptual framework (IASB, 2018) divides the qualitative characteristics of useful financial information into fundamental and enhancing characteristics. The fundamental characteristics are relevance and faithful representation; the enhancing characteristics are comparability, verifiability, timeliness, and understandability. Faithful representation means depicting the substance of the economic phenomena that the information purports to represent. Ideally, that depiction is complete, neutral, and free from error; these qualities are to be achieved to the extent possible. Verifiability means that knowledgeable and independent observers could reach consensus, although not necessarily complete agreement, that a particular depiction is a faithful representation. Understandability concerns classifying, characterizing, and presenting information clearly and concisely and presupposes users with reasonable knowledge who review the information diligently.

Applying these distinctions to mathematics permits a more precise account of what formal verification does and does not provide.

- **Verifiability:** A Lean formalization provides strong verifiability of a derivation in terms of the formalized statements, in the sense that anyone can rerun type checking and reach agreement. Here, verifiability is applied more narrowly than in the conceptual framework's definition, to derivation in terms of formalized statements.
- **Faithful representation:** Type checking does not guarantee that the formalized statements faithfully represent the substance of the intended mathematical problem. The accounting analogue is that correct arithmetic in the books does not guarantee a faithful representation of the substance of transactions.
- **Understandability:** Even when a derivation is correct and the statements are faithful, the proof need not be understandable to humans. Understandability is a

further, distinct property.

The events of 2026 illustrate the distinction among these three properties. As discussed in Section 2.3, the essay by seven formalization researchers (Avigad et al., 2026) takes the view that there is now little doubt about the correctness of the Navier–Stokes result, while describing the natural-language proof as mostly incomprehensible. This is a problem of understandability. At the same time, the argument that the Lean proof does not correspond to the natural-language proof (Section 3.2), and the observation that the natural-language and formalized versions of many released manuscripts diverge, concern faithful representation: whether what has been formalized represents the intended mathematics.

These distinctions also correspond to the model in Section 4. The status of formalization and attribution are relatively easy-to-verify measures (the difficulty of verification, ψ_2, in Section 4 is low), whereas the outward format of an exposition is a measure readily influenced by apparent compliance (the cost of apparent compliance, c_g, is low).

3.4 Global and local understanding

One approach to making proofs readable is to divide a long proof into lemmas, or modules, each small enough for humans to assimilate. Automatic blueprint generation (Zhu et al., 2026) and readability optimization (Ahuja et al., 2025) are tools for this purpose. Such tools facilitate verification and understanding of individual parts. Yet being able to assimilate every part locally is not the same as understanding why the proof as a whole works. This distinction is the present paper's formulation, but it connects with the concern for human understanding emphasized by the declaration and by Thurston (1994).

Moreover, quantities used in the objective function for improving readability, such as the rate of lemma reuse or the appropriateness of names, are themselves proxies for understanding. If such measures become targets, Goodhart's law applies at the readability stage as well. Corollary 4.1 in Section 4 formalizes the same structure as this second-stage proxy problem for the case in which a norm measures only easily verifiable components.

3.5 The framework of this paper

The framework can now be stated as follows. An AI lab's solving is a proxy measure rewarded by the market, while support for human understanding is a difficult-to-measure task (Section 4). Norms for releasing results add measures of such support, make those measures easier to verify through disclosure, and provide certification through the community channel (Sections 4 and 6). Their enforcement is limited by labs' ability to release results outside the norm (Section 4.5). An AI that provides support for human understanding also learns difficult-to-verify quality through the signal provided by human corrections (Section 5). Section 6 interprets the norms for releasing results that emerged in 2026 within this framework.

4 Model I: task allocation and crowding out through proxy measures

4.1 Question and overview of the results

Section 4 addresses research question RQ1: "How can the misalignment between the objectives of AI firms and the mathematical community be formulated as a problem of performance measurement?" In the language of performance measurement, the passage from the joint declaration quoted in Section 1 (Avila et al., 2026) can be read as saying that solved problems are a proxy measure for the objective of understanding. We formalize this reading by combining the multitask model of Holmström and Milgrom (1991), discussed in Section 3, with the concepts of congruity between a measure and an objective (Baker, 1992; Feltham and Xie, 1994; Datar et al., 2001) and distortion in performance measures (Baker, 2002).

Model I takes the reward for solving as given by the market—investors, users, and the media—rather than designed by the community. The declaration criticizes this market incentive to compete in solving problems as benchmarks. What the community can design is a measure of support for human understanding—writing up results, attribution, exposition, and formalization—and the strength of the recognition attached to it. This corresponds to release norms such as the AGMAI guidelines of September 29, 2026 (AGMAI, 2026a). The measure can also be affected by "apparent compliance" that contributes nothing to understanding. This possibility is particularly pertinent in 2026 because AI can cheaply generate polished introductions and expositions. Moreover, labs can release results without using the community channel to which the norm applies (Section 2.3), and the norm is common to labs facing different intensities of competition.

The results can be outlined as follows. Proposition 4.1 shows that when the market rewards only solving, support for human understanding is reduced if the tasks are substitutes. If the relative importance of understanding is at or above a threshold, a reward for solving reduces social surplus at every positive level. This provides a condition corresponding to the declaration's claim of harm. Proposition 4.2 shows that a norm measuring support for human understanding improves surplus, but its effect diminishes as verification becomes more difficult; if enforcement capacity is capped, the correction remains partial. Corollary 4.1 shows that a norm measuring only the more readily verifiable components can crowd out exposition that is harder to verify. Propositions 4.3–4.5 derive the limit on enforcement capacity from the lab's option to release results outside the norm. When the community's evaluation does not reach beyond the norm, the measure is noisy, and the lab is risk averse, the surplus recoverable through the norm approaches the legitimacy value of the community channel as competition intensifies (Proposition 4.3). If evaluation also reaches releases outside the norm, a strong norm with which the lab does not comply can be optimal (Proposition 4.4). With a common norm, a tension arises between the exit of the lab facing the most intense competition and the burden of the norm on the many labs that comply (Proposition 4.5). The results as the market reward increases without bound are, however, formal limits of a linear model without a floor on support for human understanding. Once a floor is imposed, a norm can no longer affect that support at a lab facing sufficiently intense competition within the available range of weights, whether or not evaluation reaches outside the norm (Section 4.5.4). Proofs and supplementary results appear in Appendices 4.A and 4.B,

details of the numerical examples in Appendix 4.C, and a list of notation in Appendix 4.D.

4.2 The model

Assumption 4.1 (Actions and costs)**.** *An AI lab (hereafter, a lab) chooses* $x = (x_1, x_2)^\top \in \mathbb{R}^2$*. Here* x_1 *is effort devoted to solving problems (solving and releasing the results), and* x_2 *is effort devoted to supporting human understanding (writing up results, searching prior literature and providing attribution, exposition, and formalization). Both are measured as deviations from the baseline levels the lab would choose in the absence of external rewards. The cost of these deviations is*

$$C(x) = \tfrac{1}{2} x^\top C x, \qquad C = \begin{pmatrix} c_1 & c_{12} \\ c_{12} & c_2 \end{pmatrix}$$

with $c_1, c_2 > 0$ *and* $\Delta \coloneqq c_1 c_2 - c_{12}^2 > 0$*. The two tasks are substitutes when* $c_{12} > 0$*.*

Assumption 4.2 (Social value)**.** *The value to the mathematical community is* $B(x) = b^\top x = b_1 x_1 + b_2 x_2$*, with* $b_1, b_2 > 0$*. We call* $\rho \coloneqq b_2/b_1$ *the relative importance of support for human understanding.*

Assumption 4.3 (Measures)**.** *The measure of solved problems,* $y_1 = x_1 + \varepsilon_1$*, is public. When a norm is in place, a measure of support for human understanding,* $y_2 = x_2 + g + \varepsilon_2$*, is constructed. Here* $g \in \mathbb{R}$ *represents apparent compliance (work that changes the measure without contributing to understanding), at a cost of* $\frac{1}{2} c_g g^2$*, where* $c_g \in (0, \infty]$*. The case* $c_g = \infty$ *means that apparent compliance is impossible; in that case,* $g \equiv 0$ *and* $1/c_g \coloneqq 0$*. The noises* $\varepsilon_1 \sim N(0, \sigma_1^2)$ *and* $\varepsilon_2 \sim N(0, \sigma_2^2)$ *are independent.*

Assumption 4.4 (Rewards and preferences)**.** *The lab's reward is* $v_1 y_1 + v_2 y_2 + (\text{constant})$*. The parameter* $v_1 > 0$ *is the strength of the reputational reward that the market attaches to solving and is taken as given. The weight* $v_2 \geq 0$ *is the strength of the recognition that the community's norm attaches to support for human understanding and is chosen by the community. The lab has CARA utility with coefficient of absolute risk aversion* $r \geq 0$*.*

Assumption 4.5 (Welfare)**.** *The community chooses* v_2 *to maximize social surplus,*

$$S \coloneqq B(x) - C(x) - \tfrac{1}{2} c_g g^2 - \tfrac{r}{2}\Big(v_1^2 \sigma_1^2 + v_2^2 \sigma_2^2\Big).$$

Rewards cancel out as transfers, whereas the risk premium and the cost of apparent compliance are social losses.

Reputation and recognition are treated as rewards with monetary value to the lab through their effects on access to funding, personnel, and users, among other channels. From the perspective of society as a whole, the gains from reputation and recognition are treated as redistribution among the parties involved (transfers). Social surplus therefore counts only the value $B(x)$ to the mathematical community, effort costs, the risk premium, and the cost of apparent compliance.

Because actions are measured as deviations from a baseline, $x_2 < 0$ means that the lab devotes less attention to writing up results or attribution than it would without rewards

(the baseline is assumed sufficiently large for the total level to remain nonnegative). The condition $c_{12} > 0$ captures a substitution relationship: the greater the pressure to release results quickly, the less time remains for writing them up and checking prior literature. Since the status of formalization can be checked by type checking, c_g is large for this component (although whether the formalized statement faithfully represents the intended problem requires a separate check). Since AI can cheaply improve the presentation of an exposition, c_g is small for that component. Hereafter, write $D \coloneqq C^{-1} = \Delta^{-1}\begin{pmatrix} c_2 & -c_{12} \\ -c_{12} & c_1 \end{pmatrix}$, $e_1 = (1,0)^\top$, $e_2 = (0,1)^\top$, and $v = (v_1, v_2)^\top$, and represent the **difficulty of verification** of measure y_2 by

$$\psi_2 \coloneqq r\sigma_2^2 + \frac{1}{c_g}.$$

The first term is the cost due to noise and risk aversion; the second is the ease of apparent compliance.

If $c_{12} > 0$, however, support for human understanding in the absence of a norm is $-(c_{12}/\Delta)v_1$ (Proposition 4.1). Thus, if the baseline is held fixed while v_1 increases without bound, the total level eventually becomes negative. The results below for $v_1 \to \infty$ (Propositions 4.3(iv), 4.4(iv), and 4.5(iii)) are stated as formal limits of the linear model without this nonnegativity constraint. Section 4.5.4 presents the implications of imposing a floor on support for human understanding.

Lemma 4.1 (Best responses and the decomposition of surplus)**.** *Under Assumptions 4.1–4.4, the lab's optimal actions are unique and given by $x^* = Dv$ and $g^* = v_2/c_g$. Moreover, the social surplus in Assumption 4.5 can be decomposed as*

$$S(v) = S^{\mathrm{FB}} - \tfrac{1}{2}(v-b)^\top D(v-b) - \tfrac{r}{2}\sigma_1^2 v_1^2 - \tfrac{\psi_2}{2}v_2^2. \tag{4.1}$$

Here $S^{\mathrm{FB}} \coloneqq \max_x\{B(x) - C(x)\} = \frac{1}{2}b^\top Db$, and the first-best action is $x^{\mathrm{FB}} = Db$.

4.3 Crowding out through rewards for solving

Proposition 4.1 (Crowding Out and Harm through a Proxy Measure). *Under Assumptions 4.1–4.5, suppose there is no norm ($v_2 = 0$).*

(i) (Crowding out) The lab's efforts are $x_1^* = (c_2/\Delta)\, v_1$ and $x_2^* = -(c_{12}/\Delta)\, v_1$. If $c_{12} > 0$, the reward for solving pushes support for human understanding below its baseline, by an amount proportional to v_1.

(ii) (The social value of rewards) Social surplus is

$$S(v_1) = A\, v_1 - \tfrac{1}{2}H_1 v_1^2, \qquad A \coloneqq \frac{c_2 b_1 - c_{12} b_2}{\Delta}, \qquad H_1 \coloneqq \frac{c_2}{\Delta} + r\sigma_1^2.$$

Hence, (a) if $A \le 0$, then $S(v_1) < S(0) = 0$ for every $v_1 > 0$; (b) if $A > 0$, then S is maximized at $v_1^\circ \coloneqq A/H_1$, and $S(v_1) < 0$ for $v_1 > 2v_1^\circ$.

(iii) (Congruity) Let $\kappa_1 \coloneqq e_1^\top Db/\sqrt{(b^\top Db)(e_1^\top De_1)}$. Then κ_1 has the same sign as A and is strictly decreasing in ρ. We have $\kappa_1 \le 0$ if and only if $c_{12} > 0$ and $\rho \ge c_2/c_{12}$. When $\kappa_1 > 0$, $S(v_1^\circ) = \kappa_1^2 \cdot c_2/(c_2 + \Delta r\sigma_1^2) \cdot S^{\mathrm{FB}} < S^{\mathrm{FB}}$.

The coefficient A is the change in social value from a unit increase in the reward for solving, evaluated at $v_1 = 0$. The reward increases solving by c_2/Δ and reduces support for human understanding by c_{12}/Δ. When $c_{12} > 0$, the condition $A \leq 0$ is equivalent to $\rho \geq c_2/c_{12}$: the relative importance of understanding is at least as large as the ratio of the own-cost coefficient of support for understanding to the strength of substitution. In this case, rewarding solving is "detrimental to the science of mathematics," in the declaration's terms. Even if $A > 0$, when $c_{12} > 0$ the optimal reward v_1° is smaller than the social value b_1. This reproduces the conclusion of Holmström and Milgrom (1991) that incentives for a measured task should be weakened when an unmeasured task is present. Since $A = x_1^{\mathrm{FB}}$, the condition $A \leq 0$ also means that solving is already socially sufficient in the absence of external rewards. This conclusion depends on defining the baseline as the allocation in the absence of external rewards.

4.4 Restoring congruity through disclosure norms and its limits

Take the market reward $v_1 > 0$ as given, and suppose that the community chooses the norm's weight $v_2 \geq 0$. Let

$$N \coloneqq c_1 b_2 - c_{12}(b_1 - v_1).$$

Since $x_2^*(v_1, 0) = -c_{12}v_1/\Delta$ and $x_2^{\mathrm{FB}} = (c_1 b_2 - c_{12} b_1)/\Delta$, we have $N = \Delta\{x_2^{\mathrm{FB}} - x_2^*(v_1, 0)\}$. Thus, $N > 0$ is equivalent to support for human understanding falling short of its first-best level in the absence of a norm. For example, $N > 0$ holds when $c_{12} \geq 0$ and $v_1 \geq b_1$.

Proposition 4.2 (Restoring Congruity through Disclosure Norms and Its Limits). *Under Assumptions 4.1–4.5, suppose $N > 0$.*

(i) (Restoration) The surplus $S(v_1, v_2)$ is strictly concave in v_2, and the optimal weight is $v_2^* = N/(c_1 + \Delta\psi_2) > 0$. The gain from the norm is $G \coloneqq S(v_1, v_2^*) - S(v_1, 0) = N^2/\{2\Delta(c_1 + \Delta\psi_2)\} > 0$. If $c_{12} > 0$, then v_2^* is strictly increasing in v_1.

(ii) (Difficulty of verification) Both v_2^* and G are strictly decreasing in ψ_2 and approach zero as $\psi_2 \to \infty$. Thus, the norm is less effective when the cost of apparent compliance c_g is lower, the noise in the measure σ_2^2 is greater (if $r > 0$), or the lab's risk aversion r is higher (if $\sigma_2^2 > 0$).

(iii) (Separation) If $\psi_2 = 0$ ($r\sigma_2^2 = 0$ and $c_g = \infty$), then $x_2^* = x_2^{\mathrm{FB}}$ and $S(v_1, v_2^*) = S^{\mathrm{FB}} - (v_1 - b_1)^2/(2c_1) - \frac{r}{2}\sigma_1^2 v_1^2$.

(iv) (A cap on enforcement capacity) Suppose the weight the norm can impose is capped at $\bar{w}_2$, where $0 \leq \bar{w}_2 < v_2^*$. The optimal weight is then $\bar{w}_2$, and the surplus shortfall is $\frac{1}{2}H_2(v_2^* - \bar{w}_2)^2$, where $H_2 \coloneqq c_1/\Delta + \psi_2$. If $c_{12} > 0$, the shortfall increases with v_1. It also decreases with ψ_2: the surplus foregone because of the cap is greater when the measure is easier to verify.

Part (iii) shows that, if the measure entails no loss from risk bearing or apparent compliance ($\psi_2 = 0$), the norm "separates" the two tasks. The remaining losses consist only of the loss from the market's misvaluation of solving and the risk premium associated with y_1. The effect of c_g in part (ii) is a specifically AI-related issue. For a measure based on the presentation of exposition, AI lowers c_g, making the norm less effective.

This conclusion applies, however, when the lab remains subject to the norm; it can be reversed when the participation constraint in Section 4.5.1 binds (Appendix 4.B.1). Taken together, parts (i) and (iv) imply that, when the tasks are substitutes, the required strength of the norm v_2^* rises as competition among AI labs strengthens the market reward v_1. If enforcement capacity $\bar{w}_2$ remains fixed, the correction increasingly falls short. Section 4.5 derives this cap from the lab's choice.

A norm measuring only the readily verifiable components The components of support for human understanding that a norm can readily measure are those that are easier to verify, such as the status of formalization and attribution. When harder-to-measure components remain—the substance of exposition for humans and support for understanding itself—the logic of Proposition 4.1 applies at a second level. Divide the activity into three tasks: x_1 is solving, x_2 is readily verifiable support, and x_3 is support that is difficult to verify. Let the cost be $\frac{1}{2}x^\top Cx$ and, to focus on the second-level problem, assume no interaction between solving and support:

$$C = \begin{pmatrix} c_1 & 0 & 0 \\ 0 & c_2 & c_{23} \\ 0 & c_{23} & c_3 \end{pmatrix}, \qquad \Delta_{23} \coloneqq c_2 c_3 - c_{23}^2 > 0.$$

Social value is $b^\top x$, with $b_1, b_2, b_3 > 0$. The measures are $y_1 = x_1 + \varepsilon_1$ (the market measure, with weight v_1) and $y_2 = x_2 + g + \varepsilon_2$ (the norm's measure, with weight $v_2 \geq 0$), while x_3 is unmeasured. All other features remain as in Assumptions 4.1–4.5.

Corollary 4.1 (A Norm Measuring Only the Readily Verifiable Components). *In the three-task setting above, the following statements hold.*

(i) The lab's efforts are $x_2^* = (c_3/\Delta_{23})v_2$ and $x_3^* = -(c_{23}/\Delta_{23})v_2$. If $c_{23} > 0$, the norm pushes the support that is difficult to verify below its baseline.

(ii) The derivative of S with respect to v_2 at $v_2 = 0$ is $M \coloneqq (c_3 b_2 - c_{23} b_3)/\Delta_{23}$. If $c_{23} b_3 \geq c_3 b_2$, then $S(v_1, v_2) < S(v_1, 0)$ for every $v_2 > 0$: a norm targeting the readily verifiable component is harmful at every positive level. In general, the optimal weight is $v_2^* = \max\{0,\ (c_3 b_2 - c_{23} b_3)/(c_3 + \Delta_{23}\psi_2)\}$.

(iii) A weight chosen while ignoring the value of x_3 (treating $b_3 = 0$ while correctly understanding the lab's cost structure and response), $v_2^{\mathrm{n}} \coloneqq c_3 b_2/(c_3 + \Delta_{23}\psi_2)$, exceeds v_2^* if $c_{23} > 0$. Moreover, $S(v_1, v_2^{\mathrm{n}}) < S(v_1, 0)$ if and only if $2c_{23} b_3 > c_3 b_2$.

If the support that is difficult to verify, x_3, is provided by a separate actor rather than the lab—for example, human mathematicians funded through an independent organization—the substitution effect c_{23} in the lab's costs does not extend to x_3, and a norm targeting the readily verifiable component does not crowd out x_3. This is the job-design logic of Holmström and Milgrom (1991): assign the harder-to-measure task to a separate actor. Such actors also overlap with those providing the "learning channels other than correction," μ_t, in Section 5.

4.5 Limits to enforcement: release outside the norm

Proposition 4.2(iv) represented enforcement capacity in reduced form as a cap $\bar{w}_2$ on the weight. In practice, a lab can release results through its own channel without using the community channel to which the norm applies. This subsection explicitly introduces that choice as an outside option and derives enforcement capacity in three stages: first, when evaluation does not reach outside the norm (Section 4.5.1); next, when evaluation partially reaches outside it (Section 4.5.2); and finally, when the norm is common to labs facing different intensities of competition (Section 4.5.3).

4.5.1 Enforcement capacity as a participation constraint

Assumption 4.6 (Choice of release channel)**.** *After the community announces v_2, and before choosing its actions, the lab chooses whether to release its results through the community channel (journals, preprints, and independent repositories to which the norm applies), denoted channel I, or through its own channel (such as GitHub, outside the norm), denoted channel O. Its reward in channel I is $v_1 y_1 + v_2 y_2 + \ell$, and its reward in channel O is $v_1 y_1$.*

- *The recognition $v_2 y_2$ attached to the norm is a relative evaluation with the baseline in Assumption 4.1 as its reference point. Support for human understanding below the baseline receives a negative assessment (criticism); support above it receives a positive assessment.*
- *The parameter $\ell \geq 0$ is the legitimacy value of releasing results through the community channel—including credibility from peer review, attribution, and access to collaborators—net of the private cost of the delay involved in using that channel. As a private value, it is excluded from social surplus.*
- *Social surplus is $S(v_1, v_2)$ in channel I and $S(v_1, 0)$ in channel O. If indifferent, the lab chooses channel I.*

Lemma 4.2 (The participation condition)**.** *Under Assumptions 4.1–4.6, the lab chooses channel I if and only if $\Phi(v_2) \geq 0$, where*

$$\Phi(v_2) \coloneqq \ell + x_2^0\, v_2 + \tfrac{1}{2} K\, v_2^2,$$
$$x_2^0 \coloneqq x_2^*(v_1, 0) = -\frac{c_{12}}{\Delta} v_1, \qquad K \coloneqq \frac{c_1}{\Delta} + \frac{1}{c_g} - r\sigma_2^2.$$

Hereafter, let $s \coloneqq -x_2^0 = c_{12} v_1 / \Delta$ and $\mathcal{V} \coloneqq \{v_2 \geq 0 : \Phi(v_2) \geq 0\}$. Since $\Phi(0) = \ell \geq 0$, we have $0 \in \mathcal{V}$. If $c_{12} > 0$, then $\Phi'(0) = x_2^0 < 0$: a weak norm initially operates as criticism of inadequate support for human understanding. The curvature K of Φ consists of the gain c_1/Δ from increasing support in response to the norm, the gain $1/c_g$ from apparent compliance, and the risk cost $r\sigma_2^2$ due to noise in the measure.

Proposition 4.3 (Enforcement Capacity as a Participation Constraint). *Under Assumptions 4.1–4.6, suppose $c_{12} > 0$, $v_1 > 0$, and $N > 0$.*

(i) (The optimal norm) The point v_2^{P} in $\mathcal{V}$ closest to v_2^* (the larger point if two points are equally close) is an optimal weight for the community. The surplus shortfall is $S(v_1, v_2^*) - S(v_1, v_2^{\mathrm{P}}) = \frac{1}{2} H_2 \operatorname{dist}(v_2^*, \mathcal{V})^2$. If a weight outside $\mathcal{V}$ is announced, the lab chooses channel O, yielding surplus $S(v_1, 0) \leq S(v_1, v_2^{\mathrm{P}})$.

(ii) (The participation set) If $K \leq 0$, then $\mathcal{V} = [0, \bar{v}]$, where $\bar{v} := 2\ell/(s + \sqrt{s^2 + 2|K|\ell})$. The bound $\bar{v}$ increases with ℓ and is nonincreasing in v_1, and $v_2^{\mathrm{P}} = \min\{v_2^*, \bar{v}\}$. Thus, the cap in Proposition 4.2(iv) is endogenously determined as $\bar{w}_2 = \bar{v}$. If $K > 0$, then $\mathcal{V}$ is either $[0, \infty)$ or of the form $[0, v_-] \cup [v_+, \infty)$, where $v_- < v_+$ are the two roots of Φ. In the latter case, only norms of intermediate strength induce exit (Proposition 4.B.1(i)).

(iii) (Selection into exit) When labs differ only in v_1 and face a common norm $v_2 > 0$, the labs choosing channel O are those with

$$v_1 > v_1^\dagger(v_2) := \frac{\Delta}{c_{12}}\Big(\frac{\ell}{v_2} + \frac{K}{2}v_2\Big).$$

These are the labs with the largest shortfalls N in support for human understanding in the absence of a norm.

(iv) (Hollowing out) For every $v_2 \in \mathcal{V}$,

$$S(v_1, v_2) - S(v_1, 0) \leq \ell + x_2^{\mathrm{FB}} v_2 - r\sigma_2^2 v_2^2,$$

with equality when $\Phi(v_2) = 0$. If $r\sigma_2^2 > 0$, the gain from the norm subject to the participation constraint is bounded above, independently of v_1, by

$$\ell + \frac{\max\{x_2^{\mathrm{FB}}, 0\}^2}{4r\sigma_2^2}.$$

As $v_1 \to \infty$, $v_2^{\mathrm{P}} \to 0$ and the gain converges to ℓ. By contrast, the unconstrained gain G diverges as $v_1 \to \infty$.

The bound $\bar{v}$ in part (ii) expresses the enforcement capacity in Proposition 4.2(iv) as a function of the value ℓ of the community channel and the market reward v_1. The participation set expands with ℓ and $1/c_g$ and contracts with v_1 and $r\sigma_2^2$ (Appendix 4.B.1). Part (iii) shows that the labs exiting a common norm are those most in need of correction through it. Part (iv) shows that, in the presence of noise and risk aversion ($r\sigma_2^2 > 0$), the surplus recoverable through the norm subject to participation has a bound independent of the market reward. Once competition becomes sufficiently intense, the norm is weakened ($v_2^{\mathrm{P}} \to 0$), and recoverable surplus approaches ℓ. These conclusions depend on placing the reference point for recognition at the baseline, on the absence of a cap on the recognition weight when $K > 0$, and on $r\sigma_2^2 > 0$ (Appendix 4.B.1).

4.5.2 Partial compliance and a cap on recognition

The community's evaluation can partially reach labs that release results outside the norm, for example through criticism of those releases. If the recognition weight is uncapped, the community may be able to announce a nominal weight that induces the lab to exit and thereby impose its desired effective weight (Appendix 4.B.2). The strength of the recognition or criticism the community can confer is, however, limited.

Assumption 4.7 (Partial compliance and a cap on recognition)**.** *Modify Assumption 4.6 as follows. (a) The recognition weight the community can announce is capped at $\bar{W} \in (0, \infty)$, so that $v_2 \in [0, \bar{W}]$. (b) The community's evaluation also applies in channel O,*

with weight θv_2, where $0 \le \theta < 1$. The reward in channel O is $v_1 y_1 + \theta v_2 y_2$. (c) Social surplus is $S(v_1, v_2)$ in channel I and $S(v_1, \theta v_2) - \zeta$ in channel O. Here $\zeta \ge 0$ is a social loss specific to channel O, such as the propagation of errors resulting from the absence of peer review. (d) If indifferent, the lab chooses the channel yielding greater surplus for the community.

Part (d) is a standard convention ensuring the existence of an optimum. The community is assumed to evaluate the lab according to its announced weight, whichever channel the lab chooses. We call the weight actually faced by the lab—v_2 in channel I and θv_2 in channel O—the **effective weight**, denoted $\tilde{v}_2$.

Lemma 4.3 (The participation set under partial compliance)**.** *Under Assumptions 4.1–4.5 and Assumption 4.6 as modified by Assumption 4.7, suppose $c_{12} > 0$ and $v_1 > 0$. The lab chooses channel I if $\Phi_\theta(v_2) > 0$ and channel O if $\Phi_\theta(v_2) < 0$, where*

$$\Phi_\theta(v_2) \coloneqq \ell - (1-\theta) s\, v_2 + \tfrac{1}{2}(1-\theta^2) K\, v_2^2 = (1-\theta)\Big\{ \ell_\theta - s\, v_2 + \tfrac{1}{2} K_\theta v_2^2 \Big\},$$

with $\ell_\theta \coloneqq \ell/(1-\theta)$ and $K_\theta \coloneqq (1+\theta)K$. The participation set $\mathcal{V}_\theta \coloneqq \{v_2 \ge 0 : \Phi_\theta(v_2) \ge 0\}$ is therefore obtained by replacing (ℓ, K) in Propositions 4.3(ii) and 4.B.1(i) with (ℓ_θ, K_θ). It can be written as $\mathcal{V}_\theta = [0, v_\theta^-] \cup [v_\theta^+, \infty)$, with the convention $[\infty, \infty) = \emptyset$. If $K \le 0$, then $v_\theta^- = 2\ell_\theta / (s + \sqrt{s^2 + 2|K_\theta|\ell_\theta})$ and $v_\theta^+ = \infty$ (see Appendix 4.B.2 for $K > 0$). The set of weights inducing exit, $\{v_2 \ge 0 : \Phi_\theta(v_2) < 0\}$, shrinks with θ in the sense of set inclusion, and v_θ^- is nondecreasing in θ (strictly increasing if $K \le 0$ and $\ell > 0$).

Proposition 4.4 (Partial Compliance and a Cap on Recognition). *Under Assumptions 4.1–4.5 and Assumption 4.6 as modified by Assumption 4.7, suppose $c_{12} > 0$, $v_1 > 0$, and $N > 0$. Denote the greatest surplus the community can attain by $\mathcal{S}_\theta^*$ (if retaining the lab in channel I and inducing it to exit yield equal surplus, the community retains it).*

(i) (Effective enforcement capacity) If $\bar{W} < v_\theta^+$ (which always holds when $K \le 0$) and $\zeta = 0$, the optimal effective weight is

$$\tilde{v}_2 = \min\{v_2^*,\ R_\theta\}, \qquad R_\theta \coloneqq \max\Big\{\min\{v_\theta^-, \bar{W}\},\ \theta\bar{W}\Big\}.$$

Thus, the cap $\bar{w}_2$ in Proposition 4.2(iv) is determined as R_θ.

(ii) (A norm that is not followed) Suppose $\bar{W} < v_\theta^+$, $v_2^* > v_\theta^-$, and $v_\theta^- < \bar{W}$. The community induces the lab to exit (because doing so yields strictly greater surplus) if and only if

$$S(v_1, \min\{v_2^*, \theta\bar{W}\}) - S(v_1, v_\theta^-) > \zeta.$$

If $\zeta = 0$, this condition is equivalent to $\theta\bar{W} > v_\theta^-$. The announced weight is then $\min\{v_2^*/\theta, \bar{W}\}$, which is $1/\theta$ times the effective weight $\min\{v_2^*, \theta\bar{W}\}$.

(iii) (The effect of θ) We have $\mathcal{S}_\theta^* \le S(v_1, \min\{v_2^*, \bar{W}\})$. The attainable surplus $\mathcal{S}_\theta^*$ is nondecreasing over the range of θ satisfying $\bar{W} < v_\theta^+$ (and over all of $[0, 1)$ if $K \le 0$).

(iv) (Intensifying competition) Fix $\zeta \geq 0$ and let $v_1 \to \infty$. The gain attainable while retaining the lab in channel I converges to $\ell/(1-\theta)$. If $\theta > 0$, then for sufficiently large v_1 the optimal weight is the cap $\bar{W}$, and the lab chooses channel O. The resulting gain, $\mathcal{S}^*_\theta - S(v_1, 0) = (N/\Delta)\theta\bar{W} - \frac{1}{2}H_2(\theta\bar{W})^2 - \zeta$, diverges, but its ratio to G approaches zero. If $\theta = 0$, the lab remains in channel I, and the gain converges to ℓ.

The quantity R_θ in part (i) represents the better of two instruments: retaining the lab through a weak norm up to the participation boundary ($\min\{v^-_\theta, \bar{W}\}$), or announcing a strong norm that the lab does not follow and imposing weight $\theta\bar{W}$ through evaluation outside the norm. In the exit-inducing solution in part (ii), the norm is written more stringently than the weight that actually takes effect. This solution is unchanged even if the community can reselect the weight after the lab chooses channel O; what is required is that releases outside the norm actually be evaluated (the credibility of θ). By contrast, in the solution that retains the lab through the weak norm v^-_θ, the community would wish to raise the weight after the lab chooses channel I. Commitment to the announced weight is therefore necessary, as it is for the solution in Proposition 4.3. Part (iv) shows that, in the linear model without a floor on support for human understanding, the hollowing out in Proposition 4.3(iv)—the gain leveling off at ℓ—is specific to $\theta = 0$. Since the effective weight cannot exceed $\theta\bar{W}$, however, the ratio of the gain to its unconstrained counterpart approaches zero even when $\theta > 0$. In the regime where the lab is retained, raising ℓ works through v^-_θ; in the regime where exit is induced, $\theta\bar{W}$ is the relevant instrument. By part (iv), stronger competition leads to the latter regime within the range in which support for human understanding does not reach its floor (Section 4.5.4 discusses the case with a floor). Appendix 4.B.2 provides the general solution when the cap exceeds the upper participation boundary, together with an example of nonmonotonicity in θ in that case.

4.5.3 A common norm for heterogeneous labs

Propositions 4.3 and 4.4 allowed the norm to be tailored to the lab's v_1. Release norms such as the AGMAI guidelines, however, are common to all labs.

Assumption 4.8 (A common norm)**.** *There are J types of labs. Labs of type $j = 1, \ldots, J$ differ only in the market reward $v_1(j)$, with $v_1(1) < \cdots < v_1(J)$. Let $\varpi_j > 0$ be the share of type j, with $\sum_j \varpi_j = 1$. Assumptions 4.1–4.5 and Assumption 4.6, as modified by Assumption 4.7, apply to each lab; ℓ, θ, $\bar{W}$, and ζ are common to all labs. The community announces a single weight $u \in [0, \bar{W}]$, and each lab chooses its channel (if indifferent, it chooses the channel yielding greater surplus for the community). The community maximizes the weighted sum of surpluses $\sum_j \varpi_j S_j$. Here S_j is the surplus for type j: $S(v_1(j), u)$ in channel I and $S(v_1(j), \theta u) - \zeta$ in channel O.*

For type j, denote the optimal weight from Proposition 4.2 by v^*_j, let $s_j \coloneqq c_{12}v_1(j)/\Delta$, write the participation condition in Lemma 4.3 as $\Phi_\theta(u; j)$, and denote the participation boundaries by $v^\pm_\theta(j)$. If $c_{12} > 0$ and $N(v_1(1)) > 0$, then $N > 0$ for every type and v^*_j is strictly increasing in j. The gain for type j, as a function of the effective weight x, is $g_j(x) \coloneqq S(v_1(j), x) - S(v_1(j), 0) = H_2(v^*_j x - \frac{1}{2}x^2)$. Denote the weighted sum of gains under norm u by $\Gamma_\theta(u) \coloneqq \sum_j \varpi_j\{S_j - S(v_1(j), 0)\}$ and its maximum by Γ^*_θ. The

operators E_ϖ and Var_ϖ denote the mean and variance with weights ϖ_j. We also define condition (T) as follows:

$$\text{(T)} \qquad K \le 0, \quad \text{or} \quad K > 0 \text{ and } \bar{W} \le \sqrt{2\ell_\theta / K_\theta}.$$

Proposition 4.5 (A Common Norm). *Under Assumption 4.8, suppose $c_{12} > 0$ and $N(v_1(1)) > 0$.*

(i) (Sorting) For every $u \in (0, \bar{W}]$, the labs choosing channel O satisfy

$$v_1(j) > v_1^\dagger(u) := \frac{\Delta}{c_{12}}\Big(\frac{\ell_\theta}{u} + \frac{K_\theta}{2}u\Big)$$

and form an upper set in v_1 (at most one type is exactly indifferent, and it lies at the boundary). Under (T), the set of labs choosing channel O expands, in the sense of set inclusion, as u increases.

(ii) (The pooling bound) For every u,

$$\Gamma_\theta(u) \le \frac{H_2}{2}\Big(\mathrm{E}_\varpi v^*\Big)^2 - \zeta\, \varpi_{\mathrm{O}}(u),$$

where $\varpi_{\mathrm{O}}(u)$ is the share of labs choosing channel O. When $\zeta = 0$, equality holds if and only if either (a) all labs choose channel I at $u = \mathrm{E}_\varpi v^*$, or (b) $\theta > 0$ and all labs choose channel O at $u = \mathrm{E}_\varpi v^*/\theta$. In particular, if $\Phi_\theta(\mathrm{E}_\varpi v^*; J) \ge 0$ and $\mathrm{E}_\varpi v^* \le \bar{W}$, then $u = \mathrm{E}_\varpi v^*$ is optimal and all labs choose channel I (if $\zeta > 0$, this is the unique optimum). Moreover, the difference between the unconstrained optimum—the sum of the gains in Proposition 4.2, $\sum_j \varpi_j G_j$—and Γ_θ^* is always at least $(H_2/2)\mathrm{Var}_\varpi(v^*)$.

(iii) (The influence of the lab facing the most intense competition) Hold all types other than J fixed and let $v_1(J) \to \infty$. If $\theta > 0$, then for sufficiently large $v_1(J)$ the optimal common norm is the cap $\bar{W}$, and type J, together with any other labs satisfying $v_1(j) > v_1^\dagger(\bar{W})$, chooses channel O. Now suppose $\theta = 0$ and $\ell > 0$, and denote by Γ_{-J}^* the maximum weighted sum of gains in the economy excluding type J, with the original weights left unchanged. For sufficiently large $v_1(J)$, if $\varpi_J(\ell + \zeta) < \Gamma_{-J}^*$, the optimal common norm is one of the optimal common norms for the economy excluding type J, and type J chooses channel O. If $\varpi_J(\ell+\zeta) > \Gamma_{-J}^*$, the optimal common norm is no greater than the participation boundary $v_0^-(J)$ $(\to 0)$ of type J, and all labs choose channel I.

(iv) (The effect of θ) The gain Γ_θ^* need not be monotone in θ. However, if an optimal common norm u^* at θ is no greater than v_j^* for every lab choosing channel O under that norm, then $\Gamma_{\theta'}^* \ge \Gamma_\theta^*$ for every $\theta' \in (\theta, 1)$.

The bound in part (ii) arises because the two-tier structure under a common norm differentiates weights in the wrong direction. Labs choosing channel I face weight u, while those choosing channel O face θu; yet it is the labs with larger v_j^* that choose channel O. Thus, the labs that need a stronger norm face a weaker effective weight. Inducing exit can therefore be useful only when participation constraints or the cap prevent the community from retaining all labs in channel I at $\mathrm{E}_\varpi v^*$. In that case, if

the optimum is interior and lies neither at a participation boundary nor at the cap, the optimal norm is a weighted average of v_j^* for labs choosing channel I (with weights ϖ_j) and the nominal targets v_j^*/θ for labs choosing channel O (with weights $\theta^2\varpi_j$); see Appendix 4.B.3. If $\theta = 0$, a lab that has exited does not affect the level of the norm. As part (iii) shows, the norm is either tailored to the economy excluding that lab or weakened almost to zero in order to retain it. If $\theta > 0$, even a lab that has exited pushes the norm upward. In the model without a floor, once competition faced by that lab becomes sufficiently intense, the norm reaches its cap, exposing other labs to exit or to the burden of an excessively strong norm. The nonmonotonicity in part (iv) arises because a higher θ makes it more difficult to allow labs for which the norm is excessive to escape its burden by moving outside it (see Appendix 4.B.3 for an example).

4.5.4 A floor on support for human understanding

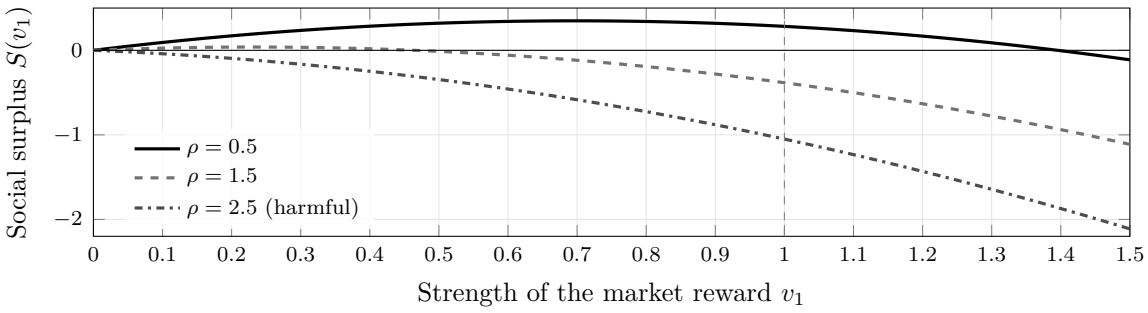
As noted in Section 4.2, Propositions 4.3(iv), 4.4(iv), and 4.5(iii) are formal limits of a linear model that does not impose nonnegativity of total levels. They should be read as describing tendencies over a range in which the baseline is sufficiently large and support for human understanding has not reached its floor. Explicitly imposing a floor changes the conclusions for sufficiently intense competition. Let the baseline be $L > 0$ and restrict the lab's choice to $x_2 \geq -L$. Under Assumption 4.7, if $v_1 \geq (c_1\bar{W} + \Delta L)/c_{12}$, support for human understanding at the lab remains at the floor $-L$, irrespective of the announced weight and the channel, and the norm merely increases apparent compliance and risk bearing. An optimal norm is then zero, and the gain from the norm is zero regardless of θ (Proposition 4.B.4). In the setting of Proposition 4.3 as well, if $K \leq 0$, the gain is zero for sufficiently large v_1. The hollowing out in Proposition 4.3(iv) thus takes an even stronger form under a floor. By contrast, the divergence of the gain in Proposition 4.4(iv), and the conclusion in Proposition 4.5(iii) for $\theta > 0$, are specific to the model without a floor. Evaluation outside the norm supplements enforcement capacity only while support for human understanding still responds to the norm. If competition intensifies beyond that range, and the cap on recognition $\bar{W}$ remains fixed, the analysis suggests two directions for a response: weakening the market reward (Proposition 4.1) and separating tasks by entrusting the support that is difficult to verify to another actor (the discussion following Corollary 4.1).

4.6 Numerical examples

Figures 1 and 2 illustrate the propositions using the baseline parameters $c_1 = c_2 = 1$, $c_{12} = 0.5$ ($\Delta = 0.75$), $b_1 = 1$, $r = 1$, and $\sigma_1^2 = 0.1$. The baseline level is assumed large enough to keep total levels nonnegative throughout the plotted range. Since support for human understanding in the absence of a norm is $-c_{12}v_1/\Delta$, the baseline L must satisfy $L \geq c_{12}v_1/\Delta$. For Figure 2(b), this requires $L \geq 26.7$ (see Section 4.5.4 for the case in which the floor binds). In Figure 1(a), the threshold beyond which rewards are harmful at every positive level is $\rho = c_2/c_{12} = 2$. Even at $\rho = 1.5$, surplus is negative when the market values solving at its social value, $v_1 = b_1 = 1$ (the optimal reward is $v_1^\circ \approx 0.23$). Figure 1(b) uses $\rho = 1.5$ and $v_1 = 1$; the cap $\bar{w}_2 = 0.6$ binds for $\psi_2 < 2$.

Figure 2 uses $b_2 = 1.5$, $c_g = 2$, and $\sigma_2^2 = 2.5$ ($\psi_2 = 3$ and $K = -2/3$). In panel (a), when $\theta = 0$ (Proposition 4.3), the participation constraint begins to bind at $v_1 \approx 0.26$, and the gain approaches $\ell = 0.1$. When $\theta > 0$, it becomes optimal to induce exit once

(a) Proposition 4.1: Rewards for solving and surplus



(b) Proposition 4.2: Gains and difficulty of verification

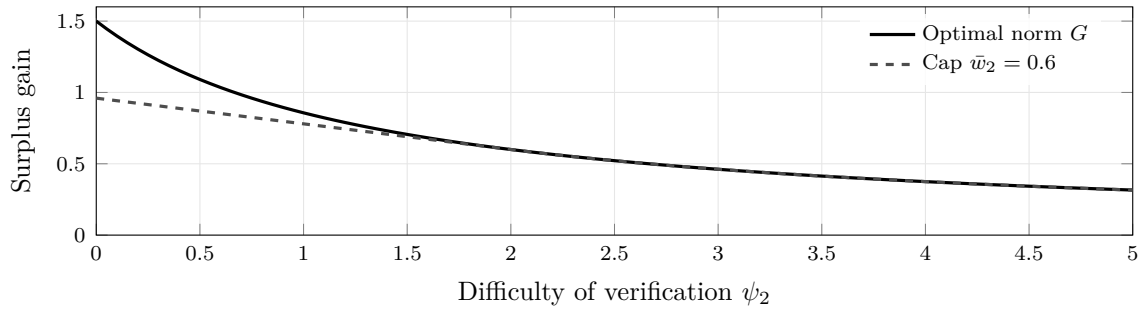


Figure 1: Numerical examples for Model I (Propositions 4.1 and 4.2). (a) When $\rho \geq c_2/c_{12} = 2$, the reward for solving reduces surplus at every positive level (the dashed line marks $v_1 = b_1$). (b) The gain shrinks as verification becomes more difficult, and the cap on enforcement capacity entails more foregone surplus for measures that are easier to verify.

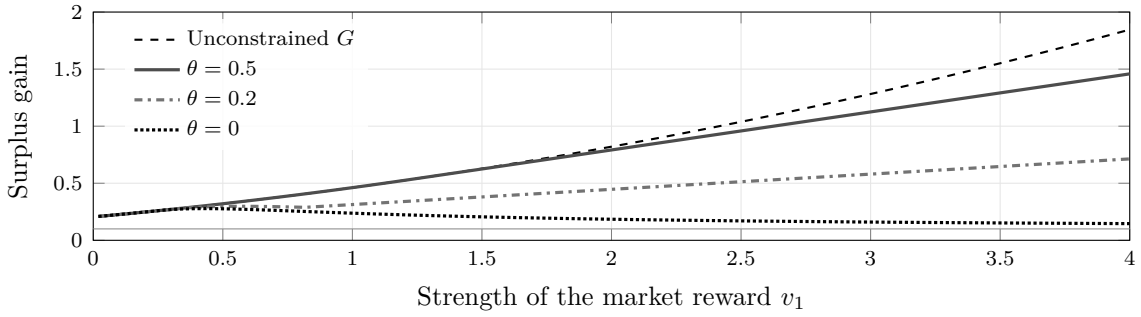
v_1 exceeds a threshold (0.818 for $\theta = 0.2$ and 0.494 for $\theta = 0.5$), after which the gain increases with v_1 (Proposition 4.4(iv)). In panel (b), when $\theta = 0$, only type 3 exits for $v_1(3) \geq 2.97$, and the norm remains constant at 0.436, tailored to types 1 and 2. When $\theta = 0.2$, after type 3 exits ($v_1(3) \geq 3.34$), the norm becomes stronger as competition faced by that type intensifies. At $v_1(3) \geq 20.8$, the norm reaches the cap and all labs exit (Proposition 4.5(iii)). Appendix 4.C provides the numerical details.

4.7 Summary and limitations

Section 4 has formalized the declaration's claim as a multitask problem in which incentives for a measured task crowd out an unmeasured task. When the tasks are substitutes, rewards for solving crowd out support for human understanding and, under certain conditions, are harmful at every positive level. A norm measuring support for human understanding restores congruity, but its effectiveness is limited by the difficulty of verification and by enforcement capacity. Enforcement capacity is determined by the lab's option to release results outside the norm. Its properties depend on how far evaluation reaches outside the norm and on the fact that the norm is common to all labs.

The limitations of Model I are discussed together in Section 8.

(a) Propositions 4.3 and 4.4: Competition and surplus gains



(b) Proposition 4.5: The lab facing the most competition and the common norm

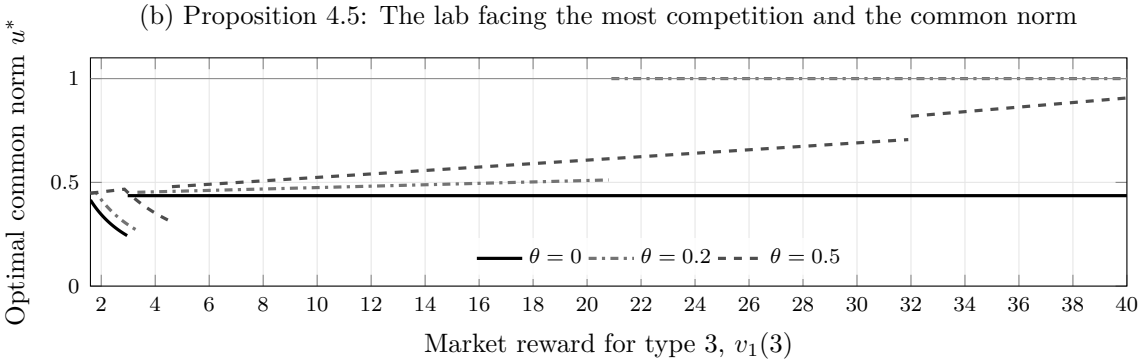


Figure 2: Numerical examples for Model I (Propositions 4.3–4.5). (a) $\ell = 0.1$, $\bar{W} = 1$, and $\zeta = 0$. When evaluation does not reach outside the norm ($\theta = 0$), the gain approaches ℓ (the thin line). When $\theta > 0$, the gain increases with v_1 after the solution enters the regime in which exit is induced. (b) A common norm for three types of labs ($v_1(1) = 0.5$ and $v_1(2) = 1.5$, with shares 0.6, 0.3, and 0.1). Here $\ell = 0.5$, $\bar{W} = 1$ (the thin line), and $\zeta = 0.05$. While type 3 is retained, the norm becomes weaker. After type 3 exits, the norm is constant if $\theta = 0$, but strengthens with competition faced by type 3 if $\theta > 0$.

5 Model II: correction-based learning and its lower bound

5.1 Research question and overview of results

Section 5 addresses research question RQ2: under what conditions does an AI that learns from human corrections converge to human mathematical taste, and where does a gap remain? As discussed in Sections 3 and 4, the quality of support for human understanding—writing up results, providing explanations, and extracting or naming lemmas—is difficult to verify. The AI therefore learns human criteria of judgment, hereafter termed mathematical taste, through human corrections to its initial output. A pair consisting of the human-corrected output and the original output can serve as material for preference learning, such as direct preference optimization (DPO; Rafailov et al., 2023), or for in-context learning. On a naive view, as use expands and training data accumulate, AI taste will converge to human taste regardless of whether corrections occur, eventually making human corrections unnecessary.

Model II starts from the observation that corrections generate learning signals. If a human makes no correction, the two outputs constituting the preference pair are identical, and the learning signal disappears. Below, we formulate this observation as an assumption about the learning mechanism: the AI updates its taste incrementally in response to current-period corrections and makes no update in periods without correction. Reusing past correction data and inferring the gap from the amount of correction lie outside this assumption (Section 5.2). Under this assumption, we obtain the following results. Proposition 5.1 shows that a positive lower bound on the gap between AI and human taste remains whenever the marginal cost of correction is positive. Corrections do not cease in finite time along the correcting path, but the gap approaches a lower bound rather than zero. Proposition 5.2 shows that the existence of a positive lower bound does not depend on the functional form of learning under the stated assumptions, although the bound weakens to $\gamma \min\{e_0, \bar{e}\}$. Proposition 5.3 shows that the gap converges to zero if and only if error reduction through other learning channels continues to accumulate. Proposition 5.4 shows that, if correction is the only opportunity to maintain human taste and taste depreciates through disuse, corrections cease in finite time and the gap becomes locked in at a higher level. Proposition 5.5 shows that corrections are socially underprovided because their benefits extend to the editor's future self and to other users. Proofs are provided in Appendix 5.A; details of the numerical illustration and a notation table (Table 11) appear in Appendix 5.B.

5.2 The model

We represent taste in one dimension. Let $e_t \geq 0$ denote the gap between the AI's initial output and human taste in period t, hereafter the taste error, and let $a_t \geq 0$ denote the amount of human correction.

Assumption 5.1 (Human Correction)**.** *In each period, taking e_t as given, the human maximizes*

$$u(a; e) \coloneqq -\frac{w}{2}(e-a)^2 - \tau a - \frac{c_0}{2}a^2 \qquad (w > 0,\ \tau \geq 0,\ c_0 > 0) \tag{5.1}$$

over $a \geq 0$.

The first term is the loss from the gap remaining after correction; the second is a linear cost per unit of correction, representing the effort of reading and checking the portions to be corrected; and the third is a convex labor cost that increases with the amount of correction. Because of the second term, the marginal cost of correction, $\tau + c_0 a$, remains at least τ even as the amount of correction approaches zero. Hereafter, we refer to τ, the marginal cost at zero correction, as the marginal cost of correction. This differs from a fixed review cost incurred once whenever a decision to correct is considered.

Assumption 5.2 (No Correction, No Signal)**.** *If $a_t = 0$, then $e_{t+1} = e_t$.*

The rationale is as follows. The DPO loss is $-\log\sigma\Big(\beta_{\mathrm{D}}[\log\frac{\pi_\vartheta(y_w)}{\pi_{\mathrm{ref}}(y_w)} - \log\frac{\pi_\vartheta(y_l)}{\pi_{\mathrm{ref}}(y_l)}]\Big)$ (Rafailov et al., 2023). If the chosen output y_w and the rejected output y_l are identical, the expression in brackets is zero regardless of the parameter ϑ. The loss is the constant $\log 2$, and its gradient is zero. Likewise, identical outputs provide no preference information in in-context learning. It is also possible to infer the gap from acceptance or approval without correction, or to update by reusing past correction data. This paper assumes, however, that such updates do not occur; we return to this restriction after Lemma 5.1.

Assumption 5.3 (Linear Learning)**.** $e_{t+1} = e_t - \lambda_t a_t$*, where* $\lambda_t \in [0, 1]$.

The interpretation is that a correction indicates the direction and magnitude of the discrepancy, and the AI moves its taste toward human taste by the fraction λ_t of that correction. Assumption 5.3 is the simplest specification satisfying Assumption 5.2; Proposition 5.2 dispenses with Assumption 5.3. Write $\omega \coloneqq w/(w + c_0) \in (0, 1)$ for the correction response coefficient, $\bar{e} \coloneqq \tau/w$, and $x^+ \coloneqq \max\{x, 0\}$.

Lemma 5.1 (Optimal Correction)**.** *Under Assumption 5.1, the optimal amount of correction is*

$$a^*(e) = \omega\,(e - \bar{e})^+ = \max\Big\{0,\ \frac{we - \tau}{w + c_0}\Big\}. \tag{5.2}$$

In particular, $a^*(e) > 0 \iff e > \bar{e}$, and, if $e > 0$, then $0 \le a^*(e) < e$.

The quantity $\bar{e}$ is the error threshold above which the human begins to correct. If the gap is at or below $\bar{e}$, the marginal effort cost τ is at least as large as the marginal benefit of correction, so the human makes no correction.

Assumptions 5.2 and 5.3 concern the learning mechanism; they are not conclusions about AI in general. For example, Lemma 5.1 implies that $e = a^*/\omega + \bar{e}$ whenever $a^* > 0$. One could therefore envisage a learning mechanism that infers the gap from the amount of correction and reaches the target in a single update ($e_{t+1} = 0$). This would violate condition (G2) in Proposition 5.2. Learning might also repeatedly reuse past correction data or draw on human preference judgments between distinct candidates without requiring edits. These possibilities can be treated as other learning channels, represented by μ_t in Proposition 5.3. Propositions 5.1 and 5.2 below concern learning mechanisms that make incremental updates in response to current-period corrections.

5.3 The lower bound on correction-based learning

Proposition 5.1 (The Lower Bound on Correction-Based Learning). *Suppose that Assumptions 5.1 and 5.3 hold, $e_0 > 0$, and the human chooses the optimal correction $a_t^* = a^*(e_t)$ from Lemma 5.1 in every period. Then the following statements hold.*

(i) (Monotonicity and non-crossing of the lower bound) The sequence e_t is nonincreasing. If $e_0 \leq \bar{e}$, then $a_t^* = 0$ and $e_t = e_0$ for every t. If $e_0 > \bar{e}$, then, for every t,

$$e_{t+1} - \bar{e} = (1 - \lambda_t\omega)(e_t - \bar{e}), \qquad e_t - \bar{e} = (e_0 - \bar{e})P_t > 0, \tag{5.3}$$

where $P_t \coloneqq \prod_{s<t}(1 - \lambda_s\omega)$, and $a_t^* = \omega(e_t - \bar{e}) > 0$. Thus, corrections do not cease in finite time.

(ii) (Limit) The limit $e_\infty \coloneqq \lim_t e_t$ exists. If $e_0 > \bar{e}$, then $e_\infty = \bar{e} + (e_0 - \bar{e})P_\infty$, and $P_\infty = 0$ if and only if $\sum_t \lambda_t = \infty$. Hence, if $\sum_t \lambda_t = \infty$, then $e_\infty = \min\{e_0, \bar{e}\}$.

(iii) (Rate) If $\lambda_t \geq \underline{\lambda} > 0$, then, for $e_0 > \bar{e}$, we have $0 < e_t - \bar{e} \leq (1 - \underline{\lambda}\omega)^t(e_0 - \bar{e})$.

(iv) (Correction amount and V) The sequence $V_t \coloneqq \frac{1}{2}e_t^2$ is nonincreasing and converges to $\frac{1}{2}e_\infty^2$. If $\tau > 0$, then $V_\infty \geq \frac{1}{2}\min\{e_0, \bar{e}\}^2 > 0$. If $\sum_t \lambda_t = \infty$, then $a_t^* \to 0$.

(v) (Fixed points and stability) Suppose that $\lambda_t > 0$ for every t. A value e is a fixed point, in the sense that $e_t = e$ implies $e_{t+1} = e$, if and only if $e \in [0, \bar{e}]$. If $\tau > 0$, then $e = 0$ is stable but not attractive. If $\tau = 0$, then $e_t \to 0 \iff \sum_t \lambda_t = \infty$.

Parts (i) and (ii) show that the human continues correcting as long as the error exceeds the threshold $\bar{e}$, and the AI learns from those corrections. As the error approaches $\bar{e}$, however, the amount of correction also approaches zero, and the learning signal diminishes. Thus, however high the learning rate, the error of an AI learning from corrections cannot fall below $\bar{e}$; if $e_0 \leq \bar{e}$, it remains at e_0. The lower bound $\bar{e} = \tau/w$ is the ratio of the cost per unit of correction to the weight on the loss from the gap. A sufficiently small gap is not worth correcting, so the human does not correct it; without correction, the AI does not learn. Parts (iv) and (v) show that a decline in the Lyapunov function $V_t = \frac{1}{2}e_t^2$ in every period need not imply convergence of the error to zero. If $\lambda_t > 0$, the set of points at which V remains constant is $[0, \bar{e}]$, rather than $\{0\}$. If $e_0 > \bar{e}$ and $\sum_t \lambda_t = \infty$, then V_t converges to $\frac{1}{2}\bar{e}^2 > 0$ (see the remark in Appendix 5.A). Table 2 contrasts this result with the naive view.

Table 2: The naive view and Proposition 5.1: the case $\tau > 0$, $e_0 > \bar{e}$, $\lambda_t > 0$ for every t, and $\sum_t \lambda_t = \infty$

Naive view	Proposition 5.1
AI taste converges to human taste ($e_t \to 0$).	$e_t \to \bar{e} = \tau/w > 0$.
The amount of correction becomes zero at some point, making human corrections unnecessary.	The amount of correction remains positive while approaching zero; corrections do not cease in finite time.
The Lyapunov function decreases in each period ($\Delta V_t < 0$), and $V_t \to 0$.	Although $\Delta V_t < 0$ in each period, $V_t \to \frac{1}{2}\bar{e}^2 > 0$ (starting in $[0, \bar{e}]$ gives $\Delta V = 0$).
$e = 0$ is globally asymptotically stable.	$e = 0$ is stable but not attractive.

The existence of a lower bound does not depend on the linearity imposed by Assumption 5.3. Assumption 5.2 and an upper bound on the size of a single learning step suffice.

Proposition 5.2 (General Existence of a Lower Bound). *Suppose that learning is given by $e_{t+1} = e_t - \Lambda_t(e_t)$, where $\Lambda_t : [0, e_0] \to [0, \infty)$ satisfies the following conditions.*

(G1) *If $e \le \bar{e}$, then $\Lambda_t(e) = 0$ (no correction, no signal; by Lemma 5.1, $a^*(e) = 0 \iff e \le \bar{e}$).*

(G2) *There exists $\gamma \in (0, 1]$ such that $\Lambda_t(e) \le (1 - \gamma)e$ for every t and e (an upper bound on the size of a single learning step).*

Then e_t is nonincreasing, has a limit e_∞, and satisfies $e_\infty \ge \gamma \min\{e_0, \bar{e}\}$. In particular, if $\tau > 0$ and $e_0 > 0$, then $e_\infty > 0$.

Corollary 5.1 (Exponential Learning). *Suppose that $e_{t+1} = e_t \exp(-k_t \varphi_t(e_t))$, where $0 \le k_t \le \bar{k}$ and $\varphi_t : [0, e_0] \to [0, 1]$ satisfies $\varphi_t(e) = 0$ whenever $e \le \bar{e}$, or equivalently $a^*(e) = 0$. Then (G1) and (G2) hold with $\gamma = e^{-\bar{k}}$, and $e_\infty \ge e^{-\bar{k}} \min\{e_0, \bar{e}\}$.*

Corollary 5.1 covers both learning signals linked to whether a correction occurs and learning signals proportional to the amount of correction. In either case, a positive lower bound on the error remains (see Appendix 5.B for numerical illustrations).

5.4 Other learning channels

Proposition 5.3 (Necessary and Sufficient Conditions for the Gap to Converge to Zero). *Suppose that learning is given by*

$$e_{t+1} = (1 - \mu_t)\, e_t - \lambda_t\, \omega\, (e_t - \bar{e})^+, \tag{5.4}$$

where $\mu_t \in [0, 1)$, $\lambda_t \in [0, 1]$, and $\mu_t + \lambda_t \omega \le 1$, and suppose that $\tau > 0$ and $e_0 > 0$. Then the following statements hold.

(i) For every t, $0 < e_{t+1} \le e_t$.

(ii) $e_t \to 0 \iff \sum_t \mu_t = \infty$.

The parameter μ_t is the proportional reduction in error through other learning channels, or the strength of signals other than current corrections. Proposition 5.3 relaxes Assumption 5.2: if $\mu_t > 0$, the gap shrinks even in periods without correction. Because μ_t can include reuse of past data, $\sum_t \mu_t = \infty$ means that error reduction independent of current corrections continues to accumulate; it does not directly mean that an infinite supply of new information is available. The naive view that learning progresses with the volume of use, independently of corrections, corresponds to setting $\lambda_t = 0$ and $\sum_t \mu_t = \infty$ in Proposition 5.3 (see the numerical illustration in Section 5.7). Thus, the naive view implicitly assumes an inexhaustible continuation of error-reducing learning independent of human corrections. Possible sources of such learning include exemplary work, expository texts, and evaluation criteria developed by the mathematical community. These are, however, products of human understanding, and their supply plausibly depends on the maintenance of human taste. Endogenizing μ_t is a task for future research; see Proposition 5.4.

Convergence of the gap to zero is distinct from operating AI without human intervention and from humans understanding its content. Even in this model, humans make no correction when the gap is at or below $\bar{e}$, although a gap remains. The lock-in described by Proposition 5.4 in the next subsection is an example of a state with a positive gap and no human intervention.

5.5 Taste depreciation and lock-in

Let $q_t \in (0,1]$ denote the acuity of human taste, and suppose that the human perceives the gap with weight wq_t. That is, replace w in (5.1) with wq_t. The variable q_t represents how strongly the human registers the gap, or sensitivity to the gap. Because errors in the direction of judgment are not modeled, this is a reduced-form representation of the acuity of taste. Applying Lemma 5.1 with w replaced by wq yields

$$a^*(e,q) = \omega(q)\,(e - \bar{e}(q))^+, \qquad \omega(q) \coloneqq \frac{wq}{wq + c_0}, \qquad \bar{e}(q) \coloneqq \frac{\tau}{wq}. \tag{5.5}$$

In this subsection, correction is the only opportunity to maintain taste; maintenance through research, education, or discussion is not considered. Taste improves through the practice of correction and depreciates through disuse:

$$e_{t+1} = e_t - \lambda_t a_t^*, \qquad q_{t+1} = (1-\delta)\,q_t + \nu\,a_t^*\,(1-q_t), \qquad a_t^* \coloneqq a^*(e_t, q_t). \tag{5.6}$$

Proposition 5.4 (Taste Depreciation and Lock-In). *Suppose that $\tau > 0$, $e_0 > 0$, $q_0 \in (0,1]$, and $\delta \in (0,1)$; that $\nu \geq 0$ satisfies $\nu\,\omega(1)\,e_0 \leq 1$; and that $\lambda_t \in [\underline{\lambda}, 1]$ with $\underline{\lambda} > 0$. Then the following statements hold for* (5.6)*.*

(i) For every t, $q_t \in (0,1]$.

(ii) There exists a finite $T \geq 0$ such that $a_t^* > 0$ for $t < T$ and $a_t^* = 0$ for $t \geq T$. Thus, corrections cease in finite time and never resume.

(iii) For $t \geq T$, $e_t = e_T$. If $T \geq 1$, then $e_T > \bar{e}(q_{T-1}) \geq \tau/w$.

(iv) For $t \geq T$, $q_t = (1-\delta)^{t-T} q_T \to 0$.

Without depreciation, as in Proposition 5.1 with $q_t \equiv 1$, corrections along the correcting path do not cease in finite time, and the gap approaches τ/w. With depreciation, corrections cease in finite time, and human taste decays toward zero. If $e_0 > \tau/(wq_0)$, or equivalently $T \geq 1$, the gap becomes locked in at a level strictly greater than τ/w. Fewer opportunities for correction weaken taste, and weaker taste raises the threshold at which correction begins. This can be read as a mathematical representation of the declaration's concern about a break in the human transmission of knowledge (Avila et al., 2026). Moreover, delegating the work of making results readable and explaining them to AI may reduce opportunities for human correction (νa_t^*), thereby eroding, through depreciation, the signal of human taste on which learning depends. This conclusion applies, however, only when correction is the sole opportunity to maintain taste. Lock-in may be avoided if activities other than correction, such as research, education, and discussion, maintain taste.

5.6 Learning externalities and underprovision of corrections

Suppose that $n \geq 1$ users use the AI's initial output in every period and that, without correction, each incurs a loss of $\frac{w}{2}e_t^2$. In each period, one editor chooses the statically optimal correction from Lemma 5.1, while the other users make no corrections. Let $\tau > 0$, let the learning rate be constant at $\lambda_t \equiv \lambda \in (0,1]$, and let the discount factor be $\beta \in (0,1)$. Define social welfare by

$$W \coloneqq \sum_{t=0}^{\infty} \beta^t \Big[-\frac{w}{2}(e_t - a_t)^2 - \tau a_t - \frac{c_0}{2}a_t^2 - (n-1)\frac{w}{2}e_t^2 \Big], \tag{5.7}$$

and write

$$\Gamma_n \coloneqq 1 + \frac{n\lambda\beta}{1-\beta}, \qquad e^{\mathrm{soc}} \coloneqq \frac{\bar{e}}{\Gamma_n} = \frac{\tau}{w\,\Gamma_n}. \tag{5.8}$$

Proposition 5.5 (Learning Externalities and Underprovision of Corrections). *Under the preceding setup, the following statements hold.*

(i) $e^{\mathrm{soc}} < \bar{e}$, *and* Γ_n *is increasing in* n, λ, *and* β.

(ii) Let $e \in (e^{\mathrm{soc}}, \bar{e}]$ *and* $e_0 = e$. *Let* W_N *denote welfare along the private path, where* $a_t = 0$ *for every* t *by Lemma 5.1 and Proposition 5.1(i). Let* $W_E(\varepsilon)$ *denote welfare under the plan* $a_0 = \varepsilon$ *and* $a_t = 0$ *for* $t \geq 1$. *Then*

$$W_E(\varepsilon) - W_N = \varepsilon\,(w\Gamma_n e - \tau) - \frac{\varepsilon^2}{2}\Big[w + c_0 + \frac{n\beta w\lambda^2}{1-\beta}\Big], \tag{5.9}$$

which is positive for $0 < \varepsilon < \min\{\bar{\varepsilon}(e),\, e\}$, *where* $\bar{\varepsilon}(e) \coloneqq 2(w\Gamma_n e - \tau)/[w + c_0 + n\beta w\lambda^2/(1-\beta)]$. *Thus, for* $e \in (e^{\mathrm{soc}}, \bar{e}]$, *the private path is not socially optimal, and corrections are underprovided.*

(iii) Let $e_0 = e \geq 0$. *The plan* $a_t = 0$ *for every* t *is socially optimal if and only if* $e \leq e^{\mathrm{soc}}$. *Here, plans are chosen from the set of all bounded nonnegative sequences* (a_t).

The private lower bound $\bar{e}$ is Γ_n times the social benchmark e^{soc}. This wedge has two sources: the myopia of the static best response, since even for $n = 1$ we have $\Gamma_1 = 1 + \lambda\beta/(1-\beta) > 1$, and the benefits accruing to the other $n-1$ users. A forward-looking editor would internalize the former but not the latter. Corrections, as data on taste, have the character of a public good.

5.7 Numerical illustration

Set $w = 1.2$, $\tau = 0.08$, and $c_0 = 0.3$, so that $\omega = 0.8$ and $\bar{e} \approx 0.0667$, and let $e_0 = 1$. Let the number of submissions in each period, representing the scale of use, be $N_t = 10 + 40/\{1 + e^{-(t-24)/8.25}\}$, and set the learning rate according to $\lambda_t\omega = 1 - \exp(-\kappa\eta N_t)$, where $\kappa = 0.032$, $\eta = 0.05$, and $\lambda_t \in [0.024, 0.096]$. The path in Proposition 5.1 is then

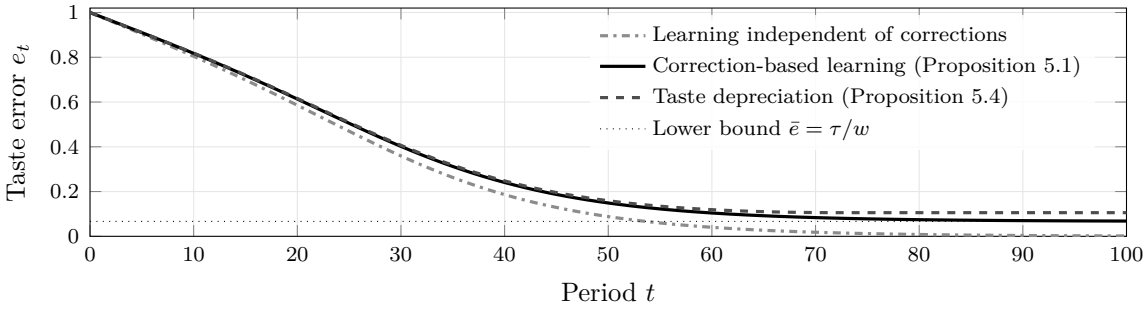
$$e_t = \bar{e} + (e_0 - \bar{e})\exp(-\kappa|\mathcal{D}_t|), \qquad |\mathcal{D}_t| \coloneqq \sum_{s<t} \eta N_s. \tag{5.10}$$

For comparison, Figure 3 also plots a path on which learning depends on the volume of use independently of corrections, $e_t = e_0 \exp(-\kappa|\mathcal{D}_t|)$, corresponding to Proposition 5.3 with $\lambda_t = 0$ and $\mu_t = 1 - \exp(-\kappa\eta N_t)$, and a path with taste depreciation, corresponding to Proposition 5.4 with $\delta = 0.02$, $\nu = 1$, and $q_0 = 1$. The path with learning independent of corrections converges to zero, whereas linking learning to corrections makes the error approach only $\bar{e}$. With depreciation, corrections cease at $t = 72$, and the error becomes locked in at approximately 0.106. Key values appear in Table 10, and numerical details are provided in Appendix 5.B.

5.8 Summary

For an AI that learns incrementally from current-period corrections, a gap determined by the ratio of correction costs to benefits remains whenever the marginal cost of correction

(a) Three Paths



(b) Proposition 5.4: A Rising Threshold and Lock-In

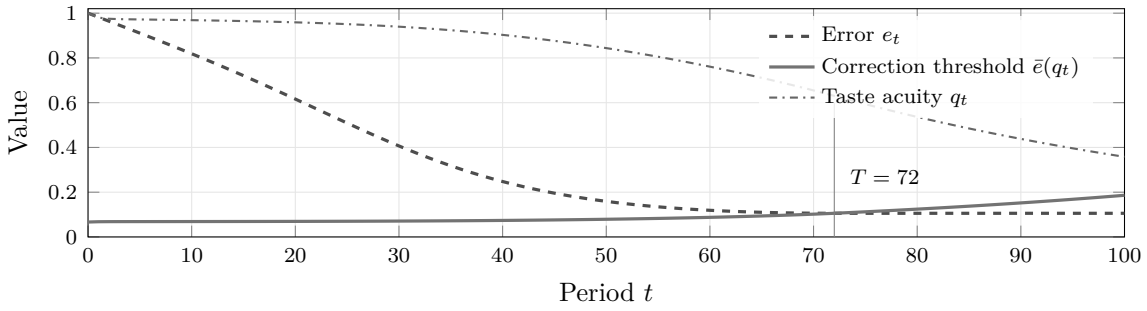


Figure 3: A numerical illustration of correction-based learning. (a) Learning independent of corrections leads the gap toward zero, whereas linking learning to corrections leads it toward $\bar{e}$ (Proposition 5.1); with taste depreciation, the gap becomes locked in above $\bar{e}$ (Proposition 5.4). (b) As taste acuity q_t declines, the correction threshold $\bar{e}(q_t) = \tau/(wq_t)$ rises. Corrections cease at $t = 72$, when the threshold exceeds the error. Thereafter, the error remains fixed while q_t continues to decay.

at zero is positive. Bringing the gap toward zero requires other learning channels, whose sources plausibly include products of human understanding. If correction is the only opportunity to maintain human taste and taste depreciates through disuse, corrections cease and the gap becomes locked in. Corrections are also socially underprovided. Section 6 uses these results to evaluate release norms, and Section 7 develops their implications for auditing, management accounting, and education. Model II is limited by its restriction to learning tied to current-period corrections, its one-dimensional representation of taste, its assumption of homogeneous human taste, and its assumption of error-free corrections (Section 8).

6 Institutional analysis: interpreting release norms

This section examines RQ3: "To what extent can the release norms that emerged in 2026 correct goal misalignment?" Using the models in Sections 4 and 5, it interprets release norms as a combination of additional measures, disclosure, and certification (Section 6.1), and examines the October 6 release and the responses to it as problems of enforcement and collective action (Section 6.2). It then considers journals, preprint servers, and formalization libraries as certification intermediaries (Section 6.3), before summarizing the conditions for correction and the limitations that remain (Section 6.4). All of the mappings and readings below are interpretations.

6.1 Release norms as measures, disclosure, and certification

Table 3 maps the main provisions of the AGMAI guidelines (AGMAI, 2026a) onto the elements of the model in Section 4. In the language of the model, these provisions fall mainly into three categories. The first is **the addition of measures**. Searches of the prior literature and attribution, readable papers with introductions and precise statements of theorems and proofs, and formalization wherever possible are components of the measure y_2 of support for human understanding. The second is **disclosure**. Records of model names, prompts, reasoning summaries, elapsed time, compute costs, the number of unsolved problems, and the selection of problems make noise and apparent compliance easier to identify, thereby reducing the difficulty of verification, ψ_2. Disclosure also facilitates the community's evaluation of results released outside the norm (represented by θ in Section 4). The third is **certification**. Depositing results in scholarly repositories not controlled by AI labs provides a channel through which third parties can certify provenance and attribution, increasing the value ℓ of releasing results through the community channel.

In addition, restraint in releasing results for promotional purposes and the cessation of testing advanced problems on proprietary models can be interpreted as provisions that weaken the market reward v_1. While norms measuring support for human understanding remain inadequate, these provisions can improve surplus in their own right (Proposition 4.1). The provision that decisions about funding conferences and expository books for human understanding should be entrusted to established nonprofit organizations corresponds to **task separation**: assigning a task that is difficult to verify to a different party (the discussion following Corollary 4.1). The Simons Foundation initiative announced on October 15, under which the foundation allocates funding that includes contributions from AI labs, can be read as an example in this direction (Simons

Table 3: Mapping the AGMAI Guidelines onto Model I: An Interpretation

Guideline provision	Place in Model I	Mechanism
Restraint in promotional releases; cessation of testing advanced problems on proprietary models	Reduces the market reward v_1	Proposition 4.1: reducing v_1 increases surplus for $v_1 > v_1^\circ$ (at all levels if $A \leq 0$)
Searches of prior literature and attribution	Relatively verifiable components of y_2	Proposition 4.2: a smaller ψ_2 makes the measure more effective, subject to the crowding out in Corollary 4.1
Readable papers (introductions, theorems, and proofs)	Components of y_2; presentation can be influenced by apparent compliance g	Proposition 4.2(ii): less effective when AI reduces c_g (the effect can reverse when the participation constraint binds)
Formalization	Relatively verifiable support x_2	Corollary 4.1: crowding out is a concern if it substitutes for substantive exposition (x_3)
Disclosure of model names, prompts, costs, and the number of unsolved problems	Disclosure of the process	Makes noise and apparent compliance easier to identify (reduces ψ_2)
Deposit in scholarly repositories not controlled by AI labs	Increases the value ℓ of the community channel	Proposition 4.B.1(iii): expands the participation set
Funding for human understanding (allocated by established nonprofit organizations)	Task separation	Discussion following Corollary 4.1 (task design)

Foundation, 2026). The provision seeking equitable access to publicly available models lies outside the models of this paper.

The model in Section 5 evaluates the same provisions from a different perspective (Table 4). Readable papers, formalization, and tools for improving readability and generating blueprints reduce the effort τ required for human correction, thereby lowering the correction-based learning floor $\bar{e} = \tau/w$. Funding for human understanding can be seen as nurturing the sources of learning channels μ_t other than correction. The emphasis on students and the human transmission of knowledge in the declaration (Avila et al., 2026) corresponds to maintaining human mathematical taste q_t.

This table permits a restatement of the value of tools for making proofs readable. Their value does not depend on their "resolving" the tensions between AI and mathematics. They are valuable because they reduce the cost τ of human correction and lower the floor of correction-based learning. By contrast, reducing the opportunities for human correction or understanding themselves—for example, by delegating even exposition and understanding to AI—can raise the floor through the depreciation mechanism in Proposition 5.4. AGMAI's position that AI labs should not lead the development of human understanding is consistent with this concern.

Table 4: Mapping Tools and Norms onto the Parameters in Section 5: An Interpretation

Parameter	Meaning	Effect on the floor	Corresponding tools and norms
Lower τ	Effort required for correction	Lowers the floor $\bar{e} = \tau/w$ (Proposition 5.1)	Readability and blueprint tools; readable papers and formalization (AGMAI)
Higher w	Private benefit of correction	Lowers $\bar{e}$ (Proposition 5.1)	Attribution and credit for human contributions (extending the emphasis on attribution in the declaration and the Leiden Declaration)
μ_t	Learning channels other than correction	The gap tends to zero if $\sum_t \mu_t = \infty$ (Proposition 5.3)	Funding for expository books and conferences (AGMAI)
$q_t,\ \delta,\ \nu$	Human mathematical taste	Depreciation leads to the cessation of correction and a locked-in gap (Proposition 5.4)	Training students and the human transmission of knowledge (declaration)
$n,\ \beta$	Sharing the benefits of correction	The private floor is Γ_n times the social benchmark (Proposition 5.5)	Subsidies for correction; attribution and compensation for correction data

6.2 The October 6 release and the community's responses: enforcement and collective action

The manuscripts released on October 6 bypassed journals, and their release was inconsistent with the guidelines' request to stop testing advanced problems on proprietary models. Some of the requested disclosures were nevertheless provided (Section 2.3). The model in Section 4 offers the following interpretation. If evaluation does not extend beyond the norm ($\theta = 0$), the exit of the lab facing the strongest competition under a common norm can be explained as a consequence of setting that norm for the majority of labs (Proposition 4.3(iii); Proposition 4.5(iii) additionally requires that this lab's weight be sufficiently small). In this case, provided that there is noise and risk aversion, the surplus that the norm can recover approaches the legitimacy value ℓ of the community channel as competition intensifies (Proposition 4.3(iv); this describes a tendency within the range in which support for human understanding does not reach its lower bound). Within this range, increasing ℓ is an objective alongside refining the provisions of the norm.

If, instead, releases following exit remain subject to evaluation through $\theta > 0$, a strong norm that the lab does not comply with can also be optimal (Proposition 4.4(ii)). This requires, however, that the participation constraint bind, that $\theta\bar{W}$ exceed the participation boundary (and, if $K > 0$, that the upper limit also lie below the upper participation boundary), and that the loss ζ specific to release without peer review not be too large. AGMAI's statement on the same day (AGMAI, 2026b) reaffirmed the guidelines while leaving the assessment of compliance to the mathematical community. AHM's statement the following day (AHM, 2026), which criticized release outside the norm, can be read as an evaluation corresponding to θ. Its call to discontinue collaboration can also be read as an attempt to reduce the value of channel O, and thus increase ℓ in relative terms, by making it harder for a lab that releases results outside the norm to obtain mathematicians' cooperation. Evaluation beyond the norm supplements

enforcement capacity, however, only while support for human understanding remains responsive to the norm. For a lab facing such intense competition that support remains at its lower bound even under a norm with weight $\bar{W}$, the norm cannot change that support within the available range of weights, regardless of θ (Section 4.5.4). Provisions that weaken the market reward and task separation are suggested as possible directions for addressing such labs. A common norm can remain effective for other labs.

The model in Section 4 nevertheless treats the community as a single designer. In practice, journals, preprint servers, formalization libraries, academic societies, an advisory group (AGMAI), and an association of researchers (AHM) act separately. The common resource to be protected here is the human attention devoted to peer review and understanding. Ostrom (1990) identified features shared by institutions that have sustained common resources over long periods, including rulemaking by users, monitoring, graduated sanctions, and, for resources that are part of larger systems, nested enterprises. The AGMAI guidelines can be read as rulemaking and AHM's call to discontinue collaboration as a sanction. arXiv's extension of submission limits to all submitters (Boboris, 2026) can likewise be read as a rule governing the use of a common resource. As discussed after Proposition 4.4 (Section 4.5.2), evaluation beyond the norm can be effective only if it is actually undertaken and is anticipated by labs when the norm is set—that is, if θ is credible. Since AHM's statement followed the release, it provides only an example of such evaluation. Some actors, such as AGMAI, leave evaluation to the community; if evaluation is dispersed among numerous actors without coordination, its credibility may be difficult to establish. The allocation of monitoring and sanctioning responsibilities among these actors therefore affects the effectiveness of the norm.

Furthermore, the guidelines constitute a common norm for labs facing different intensities of competition. When evaluation extends beyond the norm, the exit of the lab facing the strongest competition raises the norm, and the stronger norm can become excessively burdensome for the majority of labs that comply with it (Proposition 4.5(iii)). Stronger evaluation beyond the norm does not necessarily increase surplus either (Proposition 4.5(iv)). The strength of the guidelines must therefore be assessed not only in relation to the lab facing the strongest competition but also in terms of the burden on the majority of complying labs.

6.3 Journals, preprint servers, and formalization libraries as certification intermediaries

Journals, preprint servers, formalization libraries, and scholarly repositories are certification intermediaries that communicate the quality of results to third parties. How much information an intermediary discloses depends on its own incentives and on competition among intermediaries (Lizzeri, 1999). In terms of the distinctions in Section 3, these intermediaries primarily certify different things. Formalization libraries certify the correctness of derivations (verifiability), journal peer review certifies correctness and significance, and repositories certify provenance. Peer review partly addresses whether statements faithfully represent the intended problem and whether proofs are understood, but no intermediary systematically certifies these aspects as its primary function.

AGMAI's requirements for readable papers and exposition can be interpreted as an attempt to include support for human understanding within the scope of certification.

Certification based on presentation, however, readily induces apparent compliance (Lemma 4.1), which reduces its effectiveness as long as the lab remains subject to the norm (Proposition 4.2(ii)). Moreover, strengthening certification only for relatively verifiable components can crowd out the substantive content of exposition that is harder to verify (Corollary 4.1). The debate over automatic formalization of the entire mathematical literature (Avigad et al., 2026) can also be read from this perspective. Formalization greatly reduces the cost of checking derivations, but faithful representation of statements and understanding remain separate tasks.

6.4 Conditions for correction and remaining limitations

The preceding analysis yields the following account of the conditions under which release norms can correct misalignment and of what remains unresolved.

(1) Provisions that weaken the market reward, such as restraint in promotional releases, can improve surplus in their own right while norms measuring support for human understanding remain inadequate (Proposition 4.1).

(2) Norms measuring support for human understanding are more effective for components that are easier to verify, as long as labs remain subject to the norm (Proposition 4.2). Measuring only relatively verifiable components, however, can crowd out exposition that is harder to verify (Corollary 4.1). Task separation, which entrusts less verifiable support for human understanding to an independent party, may be effective.

(3) Enforcement depends on the value ℓ of the community channel and the strength $\theta\bar{W}$ of evaluation beyond the norm (Propositions 4.3 and 4.4). If $\theta > 0$, the latter is effective under intense competition within the range in which support for human understanding does not reach its lower bound. For a lab facing such intense competition that support remains at its lower bound even under a norm with weight $\bar{W}$, the norm is ineffective within the available range of weights; provisions that weaken the market reward (item (1) above) and task separation (item (2)) are suggested as possible responses. The common nature of the norm creates competing pressures on its strength (Proposition 4.5).

(4) Under the specified incremental update rule, an AI system that learns solely in response to current human correction does not eliminate a positive initial gap when the marginal cost of correction at zero is positive; the ratio of that cost to the benefit of correction determines the threshold at or below which no correction is made. Readability tools (which reduce τ) and attribution of human contributions (which increases w) can lower the corresponding learning floor (Proposition 5.1).

(5) For the gap to approach zero, error reduction through learning channels other than correction must continue to accumulate (Proposition 5.3); the sources of such learning can be viewed as products of human understanding. If correction is the only opportunity to maintain taste and taste depreciates through disuse, correction stops and the gap becomes locked in (Proposition 5.4). Mechanisms for maintaining taste therefore appear necessary.

(6) Because the benefits of correction extend to future uses and other users, correction is socially undersupplied (Proposition 5.5). Subsidies for correction, and attribution and compensation for correction data, may help address this undersupply.

Under the specified incremental update rules, learning solely from current correction

does not eliminate a positive initial gap as long as the marginal cost τ of correction at zero is positive (with linear learning, the gap does not fall below $\min\{e_0, \tau/w\}$). Within these models, learning channels other than correction are needed for the gap to approach zero, but their sources can be viewed as products of human understanding. The role of release norms can therefore be understood as establishing the conditions under which human understanding can catch up with AI-generated results, rather than removing human involvement from those results.

7 Implications for commerce

Although the analysis concerns mathematics, its structure is shared by situations in which humans verify, understand, and assume responsibility for AI-generated outputs. This section discusses implications for auditing and management accounting, skills and education, and the attribution of contributions. It then considers the place of understanding within the stages of research and the implications for the role of researchers.

7.1 Auditing and management accounting

In auditing and financial analysis, arrangements in which AI generates analyses and draft working papers for human review are becoming more prevalent. As in mathematics, these arrangements raise the problem of outputs that are "verifiable but not understood." Reperforming calculations or reconciling information may establish the correctness of a derivation without ensuring that the person responsible understands whether the assumptions faithfully represent the substance of a transaction or why the conclusion holds. In the terms of Section 3, verifiability does not guarantee faithful representation or understandability.

Proposition 5.1 also has implications for the design of these arrangements. In the model of learning in response to human review with positive τ, the ratio τ/w of the marginal cost τ of review at zero correction to the weight w on the loss from an uncorrected gap defines the threshold at or below which no correction is made. Tools that facilitate review (reducing τ) and arrangements that clarify responsibility for gaps (increasing w) both lower this threshold. Conversely, as the quality of AI's initial output improves and opportunities for correction diminish, the signal available for learning from review also becomes weaker. In performance evaluation for management accounting, measures of AI-assisted work can crowd out less verifiable judgment and the quality of explanations if they measure only readily verifiable outcomes (Corollary 4.1).

7.2 Skills and education

Proposition 5.4 shows that if correction is the only opportunity to maintain human judgment (taste), and judgment depreciates through disuse, correction stops in finite time and the AI's gap becomes locked in. In accounting education and professional training, deliberately preserving opportunities to reconsider AI outputs, rather than simply accepting them, may therefore be necessary to maintain judgment. The mechanism is a feedback process: declining judgment raises the threshold for correction, and a higher correction threshold further reduces opportunities to exercise judgment. Training

other than the correction of AI outputs, including case analysis and discussion, can also provide opportunities to maintain judgment.

7.3 Attribution of contributions

According to Proposition 5.5, correction is socially undersupplied because its benefits extend to future uses and other users. Human correction—data expressing mathematical taste—is an intangible contribution that improves AI capabilities. How to measure, attribute, and reward that contribution is a problem of the same kind as the measurement of intangible assets. The emphasis on attribution in the declaration and the Leiden Declaration (Avila et al., 2026; Leiden Declaration, 2026) indicates that the attribution of intangible contributions has also become a pressing issue within mathematics.

7.4 The stages of research and understanding

The place of the paper's conclusions becomes clearer if the research process is viewed in terms of stages: mapping reality into concepts and models ("doing physics"); moving from the creation of concepts and the formulation of problems to equivalence transformations and the compression of structure ("doing mathematics"); and developing protocols and implementing results in social and educational settings ("doing engineering" and "doing commerce"). These stages involve movement in both directions rather than a one-way progression: questions arising during implementation can, for example, lead back to problem formulation. Even if AI takes over part of the stage of "doing mathematics," it becomes difficult to examine the meaning of its results, judge their scope of application, and implement them responsibly if humans cannot understand them. Understanding can thus be seen both as a value internal to mathematics and as a condition connecting mathematical results to their use in society.

This also concerns the role of researchers. As AI assumes a greater share of solving and improving readability, the question is through which activities researchers will sustain the choice of purposes, examination of meaning, revision of evaluative criteria, and transmission to the next generation. In the language of Model II, these activities can be viewed as sources of learning channels other than correction and as opportunities for researchers to maintain their own taste. If correcting AI outputs becomes the only setting for these activities, the depreciation in Proposition 5.4 can extend to researchers themselves. The same question arises in accounting and auditing. Making explicit who needs to understand what, and to what degree, can be seen as a condition for the responsible use of AI-generated results. In this sense, the analysis reframes the declaration's concerns, from the perspective of commerce, as questions of institutional design for research and practice.

8 Conclusion

This paper has analyzed the goal misalignment between AI companies and the mathematical community that came to the fore in 2026 as a problem of performance measurement. Its answers to the three research questions are as follows.

Regarding RQ1, the formalization of goal misalignment: the declaration's statement that solving is only a proxy for understanding can be formalized as a multitask problem

in which incentives for measured tasks crowd out unmeasured tasks (Section 4). When the market rewards only solving, support for human understanding is reduced if the tasks are substitutes; if the relative importance of understanding is at least a threshold level, rewards for solving reduce social surplus at every level (Proposition 4.1). Norms that measure support for human understanding restore congruity, but, while the lab remains subject to the norm, their effectiveness diminishes as the measure becomes harder to verify (Proposition 4.2); norms that measure only relatively verifiable components can also crowd out less verifiable exposition (Corollary 4.1). Enforcement capacity is determined by the lab's option to release results outside the norm (Proposition 4.3), and its properties depend on the extent to which evaluation reaches beyond the norm (Proposition 4.4) and on the norm being common to labs (Proposition 4.5). For a lab facing such intense competition that support for human understanding remains at its lower bound even under a norm with weight $\bar{W}$, however, the norm cannot change that support within the available range of weights, whether or not evaluation extends beyond the norm (Section 4.5.4).

Regarding RQ2, learning from correction: under the specified incremental update rules, an AI system that learns solely in response to current human correction does not eliminate a positive initial gap when the marginal cost of correction at zero is positive; the ratio of that cost to the benefit of correction determines the threshold at or below which no correction is made (Propositions 5.1 and 5.2). In the extended model, the necessary and sufficient condition for the gap to tend to zero is that error reduction through learning channels other than correction continue to accumulate ($\sum_t \mu_t = \infty$; Proposition 5.3). If correction is the only opportunity to maintain taste and taste depreciates through disuse, correction stops and the gap becomes locked in (Proposition 5.4). Correction is also socially undersupplied (Proposition 5.5). Convergence of the gap to zero is distinct from operation without human intervention and from humans understanding the content.

Regarding RQ3, the corrective capacity of release norms: the release norms of 2026 can be read as additional measures of support for human understanding, disclosure of the process, and certification through the community channel (Section 6). Their corrective capacity depends on the difficulty of verifying the measure, the value of the community channel, the strength of evaluation beyond the norm, and whether learning channels other than correction and human mathematical taste are sustained. Alongside establishing norms for relatively verifiable components, it may be effective to entrust less verifiable support for human understanding to a different party and to establish mechanisms for maintaining human taste. For a lab facing such intense competition that support for human understanding remains at its lower bound even under a norm with weight $\bar{W}$, provisions that weaken the market reward and task separation are suggested as possible responses.

The paper has the following limitations. First, the models are static or employ simple dynamics; they do not examine a process in which competition among labs endogenously strengthens the market reward. Even in the analysis of a common norm, labs are represented by a finite number of types that do not interact and differ only in their market rewards. Second, Model I treats as exogenous the legitimacy value ℓ of the community channel, the benchmark for recognition, the extent θ to which evaluation reaches beyond the norm, the upper limit $\bar{W}$ on recognition weights, and the loss ζ specific to release outside the norm. It assumes that the community evaluates results according to its announced weight. The endogenous formation of ℓ and θ—through

investment in the community channel, calls to discontinue collaboration, or arrangements for evaluating releases outside the norm—and failures of commitment to the announced weight are not examined. Moreover, results obtained as the market reward increases without bound are formal limits of a linear model that imposes no lower bound on support for human understanding. The analysis with a lower bound is limited to showing the circumstances in which the norm can no longer change the relevant lab's support for human understanding. Third, social value is linear, and complementarity between solving and understanding—the role of solutions as material for understanding—is omitted. Introducing complementarity could change the threshold in Proposition 4.1. Nor are the effects of task separation, in which less verifiable support is entrusted to a different party, formally analyzed. Fourth, the community is treated as a single designer, and collective-action problems among journals, preprint servers, formalization libraries, and academic societies are discussed only qualitatively in Section 6. There is also no equation linking support for human understanding in Model I to the error in taste in Model II; the two mechanisms are analyzed separately. Fifth, Model I depends on LEN assumptions (linear rewards, CARA utility, and normally distributed noise), while Model II depends on one-dimensional taste and a learning rule that updates incrementally in response to current correction. Opportunities to maintain taste are also restricted to correction. Nonlinearity in reputation, heterogeneous tastes, and errors in correction are not examined. Neither the endogenous supply of learning channels other than correction nor the dynamics of correction by a forward-looking editor are considered. Sixth, the factual account is current as of the information cutoff specified in Section 2; assessments of ongoing events, particularly claims still under verification, may change. These extensions, together with empirical calibration using records of correction and other data, are tasks for future research.

Note on the use of AI

In preparing the Japanese manuscript, the author used conversations with Google Gemini (3.1 Pro and 4 Argon) in September and October 2026 to organize the initial ideas. In October 2026, Anthropic's Claude (Claude Opus 5.5) was used to examine prior discussions and related materials, consider the structure of the paper, formulate the propositions in Sections 4 and 5 and prepare preliminary proofs, calculate numerical examples, draft the manuscript (including the English abstract), and cross-check proofs, numerical values, and citations. AI outputs were not treated as independent verification of proofs, facts, or interpretations. The author checked all propositions, proofs, numerical values, citations, and factual statements adopted in the Japanese manuscript and assumes responsibility for the content of the final manuscript.

In October 2026, OpenAI ChatGPT was additionally used for source checking, editorial review, English translation, and typesetting of this English working paper. These uses likewise do not constitute independent verification of the paper's proofs, facts, or interpretations.

References

Advisory Group on Mathematics and Artificial Intelligence (AGMAI). (2026a). Responsible release of AI-generated mathematics. September 29, 2026. https://agmai.org/general-sep29/

Advisory Group on Mathematics and Artificial Intelligence (AGMAI). (2026b). On OpenAI's release of mathematical results. October 6, 2026. https://agmai.org/#openai-release

Advisory Group on Mathematics and Artificial Intelligence (AGMAI). (2026c). Advisory Group on Mathematics and Artificial Intelligence (website). Retrieved October 18, 2026, from https://agmai.org/

Ahuja, R., Avigad, J., Tetali, P., & Welleck, S. (2025). ImProver: Agent-based automated proof optimization. International Conference on Learning Representations (ICLR 2025).

Association for Human Mathematics (AHM). (2026). AHM statement on OpenAI's October 6 release of mathematical documents. October 7, 2026. https://www.ahmath.org/statements (reposted on T. Tao's blog: https://terrytao.wordpress.com/2026/10/07/ahm-statement-on-openais-october-6-release-of-mathematical-documents/)

Avigad, J., Buzzard, K., Commelin, J., Degenne, R., Morrison, K., Nash, O., & Topaz, A. (2026). How to formalize all of mathematics. Xena Project (blog), October 16, 2026. https://xenaproject.wordpress.com/2026/10/16/how-to-formalize-all-of-mathematics/

Avila, A., Bhargava, M., Birkar, C., Deligne, P., Deng, Y., ..., Tao, T., ..., & Zelmanov, E. (2026). A severe misalignment of AI in mathematics. September 11, 2026. https://mathandai.org (Zenodo: https://doi.org/10.5281/zenodo.22737750)

Baker, G. P. (1992). Incentive contracts and performance measurement. Journal of Political Economy, 100(3), 598–614.

Baker, G. P. (2002). Distortion and risk in optimal incentive contracts. Journal of Human Resources, 37(4), 728–751.

Bastounis, A., Circelli, F., & Hansen, A. C. (2026). Navier–Stokes lost in translation: Why Lean verification of AI autoformalisation does not guarantee correct natural language proofs. arXiv:2610.08144. https://arxiv.org/abs/2610.08144

Boboris, K. (2026). Fair moderation, equitable access, and AI: arXiv's updated rate limit policy. arXiv blog, October 1, 2026. https://blog.arxiv.org/2026/10/01/updated-rate-limit-policy/

Clay Mathematics Institute. (2026). Navier–Stokes announcement. September 11, 2026. https://www.claymath.org/news/navier-stokes-announcement/

Datar, S., Kulp, S. C., & Lambert, R. A. (2001). Balancing performance measures. Journal of Accounting Research, 39(1), 75–92.

DeMarzo, P. M., Fishman, M. J., & Hagerty, K. M. (2005). Self-regulation and government oversight. Review of Economic Studies, 72(3), 687–706.

De Millo, R. A., Lipton, R. J., & Perlis, A. J. (1979). Social processes and proofs of theorems and programs. Communications of the ACM, 22(5), 271–280.

Feltham, G. A., & Xie, J. (1994). Performance measure congruity and diversity in multi-task principal/agent relations. The Accounting Review, 69(3), 429–453.

Frey, B. S., & Jegen, R. (2001). Motivation crowding theory. Journal of Economic Surveys, 15(5), 589–611.

Holmström, B., & Milgrom, P. (1991). Multitask principal–agent analyses: Incentive contracts, asset ownership, and job design. Journal of Law, Economics, & Organization, 7(Special Issue), 24–52.

International Accounting Standards Board (IASB). (2018). Conceptual framework for financial reporting. IFRS Foundation.

Kerr, S. (1975). On the folly of rewarding A, while hoping for B. Academy of Management Journal, 18(4), 769–783.

LaSalle, J. P. (1976). The stability of dynamical systems. Society for Industrial and Applied Mathematics.

Leiden Declaration on Artificial Intelligence and Mathematics. (2026). June 2, 2026. https://leidendeclaration.ai/ (Zenodo: https://doi.org/10.5281/zenodo.20302944)

Lerner, J., & Tirole, J. (2006). A model of forum shopping. American Economic Review, 96(4), 1091–1113.

Lizzeri, A. (1999). Information revelation and certification intermediaries. RAND Journal of Economics, 30(2), 214–231.

Ma, J. W. (2026). Toward an engineering of science: Rebalancing generation and verification in the age of AI. arXiv:2605.10425. https://arxiv.org/abs/2605.10425

OpenAI. (2026a). On the Navier–Stokes Millennium Prize Problem. September 8, 2026 (updated September 10, 2026). https://openai.com/index/navier-stokes-solution/

OpenAI. (2026b). Advisory Group on Mathematics and Artificial Intelligence. September 21, 2026. https://openai.com/index/advisory-group-on-mathematics-and-ai/

OpenAI. (2026c). Sharing AI progress in mathematics. October 6, 2026 (updated October 15, 2026). https://openai.com/index/sharing-ai-progress-in-mathematics/

OpenAI. (2026d). openai/math (GitHub repository: README and history). Retrieved October 18, 2026, from https://github.com/openai/math

Ostrom, E. (1990). Governing the commons: The evolution of institutions for collective action. Cambridge University Press.

Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C. D., & Finn, C. (2023). Direct preference optimization: Your language model is secretly a reward model. Advances in Neural Information Processing Systems, 36.

Reichenbach, H. (1938). Experience and prediction: An analysis of the foundations and the structure of knowledge. University of Chicago Press.

Simons Foundation. (2026). Simons Foundation launches $15M effort to help mathematicians navigate AI era. October 15, 2026. https://www.simonsfoundation.org/news/simons-foundation-launches-15m-effort-to-help-mathematicians-navigate-ai-era/

Strathern, M. (1997). 'Improving ratings': Audit in the British University system. European Review, 5(3), 305–321.

Thurston, W. P. (1994). On proof and progress in mathematics. Bulletin of the American Mathematical Society, 30(2), 161–177.

Zhu, T., Monticone, P., Welleck, S., & Avigad, J. (2026). LeanArchitect: Automating blueprint generation for humans and AI. In E. Komendantskaya & T. Nipkow (Eds.), 17th International Conference on Interactive Theorem Proving (ITP 2026), LIPIcs 382, 25:1–25:16. https://doi.org/10.4230/LIPIcs.ITP.2026.25

Appendix 4.A Proofs for Section 4

4.A.1 Lemma 4.1 and preliminaries

Proof of Lemma 4.1. The lab's certainty equivalent is $v_1x_1+v_2(x_2+g)-C(x)-\frac{1}{2}c_g g^2-\frac{r}{2}(v_1^2\sigma_1^2+v_2^2\sigma_2^2)$ + (constant). This expression is strictly concave in (x,g) because C is positive definite and $c_g>0$. Hence the first-order conditions $Cx=v$ and $c_g g=v_2$ identify the unique maximizer. Next, $B(x^*)-C(x^*)=b^\top Dv-\frac{1}{2}v^\top Dv=\frac{1}{2}b^\top Db-\frac{1}{2}(v-b)^\top D(v-b)$. Also, $\frac{1}{2}c_g(g^*)^2=v_2^2/(2c_g)$. Combining these expressions with the risk premium yields (4.1). The maximizer of $B(x)-C(x)$ is $x^{\mathrm{FB}}=Db$, the solution to $Cx=b$. □

When $c_g=\infty$, the same argument applies with $g\equiv 0$, and Lemma 4.1 holds with $\psi_2=r\sigma_2^2$. The same observation applies to the following lemma.

Lemma 4.A.1 (Marginal social value of reward weights). *S is strictly concave in v, with $\partial S/\partial v_1=e_1^\top D(b-v)-r\sigma_1^2v_1$ and $\partial S/\partial v_2=e_2^\top D(b-v)-\psi_2v_2$.*

Proof. Differentiate (4.1). The Hessian is $-(D+\mathrm{diag}(r\sigma_1^2,\psi_2))$, which is negative definite. □

The first term values the change in effort induced by a weight, De_j, at the difference $b-v$ between social value and private rewards. The second term represents the costs of risk and apparent compliance.

4.A.2 Proof of Proposition 4.1

(i) Set $v=(v_1,0)^\top$ in Lemma 4.1.

(ii) Substituting $v=v_1e_1$ into (4.1) gives $S(v_1)=v_1e_1^\top Db-\frac{1}{2}v_1^2e_1^\top De_1-\frac{r}{2}\sigma_1^2v_1^2$. The claimed form follows from $e_1^\top Db=(c_2b_1-c_{12}b_2)/\Delta=A$ and $e_1^\top De_1=c_2/\Delta$. Since $S(v_1)=v_1(A-\frac{1}{2}H_1v_1)$ and $H_1>0$, (a) and (b) follow immediately.

(iii) Since $A=e_1^\top Db=\kappa_1\sqrt{(b^\top Db)(e_1^\top De_1)}$ and the expression under the square root is positive, κ_1 and A have the same sign. Writing $b=b_1(1,\rho)^\top$, we find that κ_1 depends only on ρ, with

$$\kappa_1(\rho)=\frac{c_2-c_{12}\rho}{\sqrt{c_2\,q(\rho)}},\qquad q(\rho)\coloneqq c_2-2c_{12}\rho+c_1\rho^2=\Delta\,(1,\rho)D(1,\rho)^\top>0.$$

Differentiation gives $\frac{d}{d\rho}\Big[(c_2-c_{12}\rho)q^{-1/2}\Big]=q^{-3/2}\Big[-c_{12}q-\frac{1}{2}(c_2-c_{12}\rho)q'\Big]=-\Delta\rho\,q^{-3/2}$, so $\kappa_1'(\rho)=-\Delta\rho/(\sqrt{c_2}\,q(\rho)^{3/2})<0$ for $\rho>0$. We have $\kappa_1\le 0\iff c_2\le c_{12}\rho$; since $c_2,\rho>0$, this is equivalent to $c_{12}>0$ and $\rho\ge c_2/c_{12}$. Finally, when $A>0$, $S(v_1^\circ)=A^2/(2H_1)=\kappa_1^2(b^\top Db)(e_1^\top De_1)/\{2(e_1^\top De_1+r\sigma_1^2)\}=\kappa_1^2S^{\mathrm{FB}}\cdot c_2/(c_2+\Delta r\sigma_1^2)$. Because $b_2>0$, b is not parallel to e_1, so the Cauchy–Schwarz inequality for the inner product $\langle u,w\rangle=u^\top Dw$ gives $\kappa_1^2<1$. Since $S^{\mathrm{FB}}=\frac{1}{2}b^\top Db>0$ and the coefficient $c_2/(c_2+\Delta r\sigma_1^2)$ is at most one, $S(v_1^\circ)<S^{\mathrm{FB}}$. □

Remark (Magnitude of congruity). The expression in the proof of (iii) gives $1-\kappa_1^2=\Delta\rho^2/(c_2\,q(\rho))$. When $\kappa_1>0$ and $r\sigma_1^2=0$, κ_1^2 is the fraction of first-best surplus attainable using only the solving measure. If $\kappa_1\le 0$, the best attainable outcome is $S=0$ at $v_1=0$. A larger relative importance of understanding, ρ, reduces this fraction within the range $\kappa_1>0$.

4.A.3 Proof of Proposition 4.2

(i) Lemma 4.A.1 and $e_2^\top D = \Delta^{-1}(-c_{12}, c_1)$ give

$$\frac{\partial S}{\partial v_2} = \frac{c_1(b_2 - v_2) - c_{12}(b_1 - v_1)}{\Delta} - \psi_2 v_2, \qquad \frac{\partial^2 S}{\partial v_2^2} = -\Big(\frac{c_1}{\Delta} + \psi_2\Big) < 0.$$

The first-order condition yields $v_2^* = N/(c_1 + \Delta\psi_2)$. For a quadratic function in v_2, $f(v_2) = f(0) + f'(0)v_2 - \frac{1}{2}Hv_2^2$, the difference between its maximum and $f(0)$ is $f'(0)^2/(2H)$. Substitution of $f'(0) = N/\Delta$ and $H = (c_1 + \Delta\psi_2)/\Delta$ gives $G = N^2/\{2\Delta(c_1 + \Delta\psi_2)\}$. Since $N > 0$, $v_2^* > 0$ and $G > 0$. Also, $\partial v_2^*/\partial v_1 = c_{12}/(c_1 + \Delta\psi_2)$. Below, we repeatedly use the representations $S(v_1, x) = S(v_1, v_2^*) - \frac{1}{2}H_2(x - v_2^*)^2$ and $S(v_1, x) - S(v_1, 0) = (N/\Delta)x - \frac{1}{2}H_2x^2$, where $H_2 = c_1/\Delta + \psi_2$.

(ii) This follows directly from the expressions in (i). The quantity $\psi_2 = r\sigma_2^2 + 1/c_g$ is increasing in r and σ_2^2 (strictly so when the other factor is positive) and strictly decreasing in c_g.

(iii) If $\psi_2 = 0$, then $v_2^* = N/c_1 = b_2 + (c_{12}/c_1)(v_1 - b_1)$. Thus $x_2^* = e_2^\top Dv^* = (-c_{12}v_1 + c_1v_2^*)/\Delta = (c_1b_2 - c_{12}b_1)/\Delta = x_2^{\mathrm{FB}}$. Moreover, $v^* - b = (v_1 - b_1)(1, c_{12}/c_1)^\top$, so

$$\tfrac{1}{2}(v^* - b)^\top D(v^* - b) = \frac{(v_1 - b_1)^2}{2\Delta}\Big(c_2 - \frac{2c_{12}^2}{c_1} + \frac{c_{12}^2}{c_1}\Big) = \frac{(v_1 - b_1)^2}{2c_1}.$$

The claimed expression follows from (4.1); for $c_g = \infty$, use the observation in Appendix 4.A.1.

(iv) Since $S(v_1, \cdot)$ is a strictly concave quadratic function and is increasing on $[0, v_2^*]$, its maximizer on $[0, \bar{w}_2]$ is $\bar{w}_2$, and the shortfall F equals $\frac{1}{2}H_2(v_2^* - \bar{w}_2)^2$. Neither H_2 nor $\bar{w}_2$ depends on v_1, and (i) implies that v_2^* increases with v_1. Hence F also increases. Next, letting $a \coloneqq N/\Delta$, we have $v_2^* = a/H_2$ and $F = (a - H_2\bar{w}_2)^2/(2H_2)$, so

$$\frac{dF}{dH_2} = -\frac{(a - H_2\bar{w}_2)(a + H_2\bar{w}_2)}{2H_2^2} = -\frac{1}{2}\Big(v_2^{*2} - \bar{w}_2^2\Big) < 0,$$

because $v_2^* > \bar{w}_2 \geq 0$. Since $dH_2/d\psi_2 = 1$, F decreases with ψ_2. □

Remark. (1) (Sign of N.) If support for understanding is already at or above its first-best level without a norm ($N \leq 0$; when $c_{12} > 0$, this means $b_1 - v_1 \geq c_1b_2/c_{12}$), then $\partial S/\partial v_2 = N/\Delta \leq 0$ at $v_2 = 0$, so the optimal weight is $v_2 = 0$. Also, if $c_{12} \geq 0$ and $x_2^{\mathrm{FB}} \geq 0$ (that is, $c_1b_2 \geq c_{12}b_1$), then $N > 0$ for every $v_1 > 0$.

(2) (Identity for the gain.) The expressions in (i) imply the general identity $G = v_2^* \cdot N/(2\Delta)$.

(3) (If the market reward could be chosen.) If the market reward could also be chosen, the desirable v_1 with an optimally chosen norm could exceed its value v_1° without a norm. For example, if $\psi_2 = 0$, maximizing the expression in (iii) over v_1 within the range $N > 0$ gives $v_1 = b_1/(1 + c_1r\sigma_1^2)$. The implication of Proposition 4.1 that incentives for solving should be weakened depends on the absence of a measure of support for understanding.

4.A.4 Proof of Corollary 4.1

The arguments establishing Lemma 4.1 and Lemma 4.A.1 extend directly to three tasks. The matrix $D = C^{-1}$ is block diagonal, with $(2,3)$ block $\Delta_{23}^{-1}\begin{pmatrix} c_3 & -c_{23} \\ -c_{23} & c_2 \end{pmatrix}$.

(i) This follows from $x^* = Dv$ and $v = (v_1, v_2, 0)^\top$.

(ii) We have $\partial S/\partial v_2 = e_2^\top D(b - v) - \psi_2 v_2 = \{c_3(b_2 - v_2) - c_{23}b_3\}/\Delta_{23} - \psi_2 v_2$. Here $(b - v)_3 = b_3$, and the block structure eliminates v_1 from this expression. Its value at $v_2 = 0$ is therefore M. Since $S(v_1, \cdot)$ is a quadratic function with second derivative $-(c_3/\Delta_{23} + \psi_2) < 0$, $S(v_1, v_2) - S(v_1, 0) = Mv_2 - \frac{1}{2}(c_3/\Delta_{23} + \psi_2)v_2^2$. If $M \le 0$, this expression is negative for every $v_2 > 0$, so the optimal weight over $v_2 \ge 0$ is zero. If $M > 0$, the first-order condition yields v_2^*.

(iii) When $c_{23} > 0$, $M > 0$ implies $v_2^{\mathrm{n}} - v_2^* = c_{23}b_3/(c_3 + \Delta_{23}\psi_2) > 0$, whereas $M \le 0$ implies $v_2^* = 0 < v_2^{\mathrm{n}}$. Next, writing $H \coloneqq c_3/\Delta_{23} + \psi_2$, we have $v_2^{\mathrm{n}} = c_3 b_2/(\Delta_{23}H)$, so the expression in (ii) gives

$$S(v_1, v_2^{\mathrm{n}}) - S(v_1, 0) = v_2^{\mathrm{n}}\Big(M - \tfrac{1}{2}Hv_2^{\mathrm{n}}\Big) = \frac{v_2^{\mathrm{n}}}{\Delta_{23}}\Big(\tfrac{1}{2}c_3 b_2 - c_{23}b_3\Big).$$

Since $v_2^{\mathrm{n}} > 0$, this is negative if and only if $2c_{23}b_3 > c_3 b_2$. □

4.A.5 Proofs of Lemma 4.2 and Proposition 4.3

Proof of Lemma 4.2. As in the proof of Lemma 4.1, the certainty equivalent in channel I is $\frac{1}{2}v^\top Dv + v_2^2/(2c_g) - \frac{r}{2}(\sigma_1^2 v_1^2 + \sigma_2^2 v_2^2) + \ell$, and that in channel O is $\frac{1}{2}D_{11}v_1^2 - \frac{r}{2}\sigma_1^2 v_1^2$, with common constant terms. Their difference is

$$D_{12}v_1 v_2 + \tfrac{1}{2}D_{22}v_2^2 + \frac{v_2^2}{2c_g} - \tfrac{r}{2}\sigma_2^2 v_2^2 + \ell.$$

Substituting $D_{12} = -c_{12}/\Delta$ and $D_{22} = c_1/\Delta$ yields $\Phi(v_2)$. □

Proof of Proposition 4.3 (i) If $v_2 \in \mathcal{V}$, the lab chooses channel I and surplus is $S(v_1, v_2)$. If $v_2 \notin \mathcal{V}$, it chooses channel O and surplus is $S(v_1, 0)$. Since $0 \in \mathcal{V}$, the maximum surplus attainable by the community is $\max_{v \in \mathcal{V}} S(v_1, v)$. Because $S(v_1, v) = S(v_1, v_2^*) - \frac{1}{2}H_2(v - v_2^*)^2$, a maximizer is a point in $\mathcal{V}$ closest to v_2^*. Such a point exists because $\mathcal{V}$ is closed and $\mathcal{V} \cap [0, 2v_2^*]$ is compact and nonempty. The shortfall formula follows from the same representation. Also, $|v_2^{\mathrm{P}} - v_2^*| \le |0 - v_2^*|$ implies $S(v_1, 0) \le S(v_1, v_2^{\mathrm{P}})$.

(ii) The function Φ is quadratic, with $\Phi(0) = \ell \ge 0$ and $\Phi'(0) = -s < 0$. If $K \le 0$, then $\Phi'(v) = -s + Kv < 0$ for $v \ge 0$, so Φ is strictly decreasing and tends to $-\infty$. Its unique nonnegative root is $(\sqrt{s^2 + 2|K|\ell} - s)/|K| = 2\ell/(s + \sqrt{s^2 + 2|K|\ell})$, the solution of $\frac{1}{2}|K|v^2 + sv - \ell = 0$, if $K < 0$, and ℓ/s if $K = 0$. In either case, this root equals $\bar{v}$. The expression for $\bar{v}$ is increasing in ℓ and nonincreasing in s (strictly decreasing when $\ell > 0$), while s increases with v_1. When $\mathcal{V} = [0, \bar{v}]$, its closest point to v_2^* is $\min\{v_2^*, \bar{v}\}$. If $K > 0$, Φ is strictly convex and attains its minimum $\ell - s^2/(2K)$ at $v_m \coloneqq s/K$. If the minimum is nonnegative, $\mathcal{V} = [0, \infty)$. Otherwise, writing the two roots as $v_\pm = (s \pm \sqrt{s^2 - 2K\ell})/K$, we have $0 \le v_- < v_m < v_+$, and $\Phi < 0$ holds precisely on (v_-, v_+).

(iii) For $v_2 > 0$, $\Phi(v_2) = \ell + \frac{1}{2}Kv_2^2 - (c_{12}/\Delta)v_1 v_2$ is strictly decreasing in v_1. Solve $\Phi < 0$ for v_1. The quantity $N(v_1) = c_1 b_2 - c_{12}b_1 + c_{12}v_1$ increases with v_1.

(iv) Since $N/\Delta = x_2^{\mathrm{FB}} + s$ and $H_2 - K = 2r\sigma_2^2$,

$$\begin{aligned} S(v_1, v) - S(v_1, 0) &= (N/\Delta)v - \tfrac{1}{2}H_2 v^2 = x_2^{\mathrm{FB}}v + sv - \tfrac{1}{2}Kv^2 - r\sigma_2^2 v^2 \\ &= \ell + x_2^{\mathrm{FB}}v - r\sigma_2^2 v^2 - \Phi(v). \end{aligned}$$

For $v \in \mathcal{V}$, $\Phi(v) \geq 0$, which gives the inequality, with equality when $\Phi = 0$. If $r\sigma_2^2 > 0$, the maximum of the quadratic upper bound over $v \geq 0$ is $\ell + \max\{x_2^{\mathrm{FB}}, 0\}^2/(4r\sigma_2^2)$.

Now let $v_1 \to \infty$, equivalently $s \to \infty$. Then $v_2^* = \Delta(x_2^{\mathrm{FB}} + s)/(c_1 + \Delta\psi_2) \to \infty$ and $G = N^2/\{2\Delta(c_1 + \Delta\psi_2)\} \to \infty$. If $K \leq 0$, $\bar{v} \to 0$, so for sufficiently large v_1, $v_2^{\mathrm{P}} = \bar{v} \to 0$. If $K > 0$, then $\ell < s^2/(2K)$ for sufficiently large v_1, and $v_- = 2\ell/(s + \sqrt{s^2 - 2K\ell}) \to 0$. Moreover,

$$v_2^* - \frac{s}{K} = \frac{\Delta x_2^{\mathrm{FB}}}{c_1 + \Delta\psi_2} + s\Big(\frac{\Delta}{c_1 + \Delta\psi_2} - \frac{1}{K}\Big),$$

where the term in parentheses is negative because $c_1 + \Delta\psi_2 - \Delta K = 2\Delta r\sigma_2^2 > 0$. Thus, for sufficiently large v_1, $v_- < v_2^* < s/K$, and Proposition 4.B.1(i) gives $v_2^{\mathrm{P}} = v_- \to 0$. In either case, for sufficiently large v_1, v_2^{P} is a root of Φ (either $\bar{v}$ or v_-), so $\Phi(v_2^{\mathrm{P}}) = 0$. The identity in the first part of the proof therefore implies that the gain $\ell + x_2^{\mathrm{FB}} v_2^{\mathrm{P}} - r\sigma_2^2 (v_2^{\mathrm{P}})^2$ converges to ℓ. □

4.A.6 Proofs of Lemma 4.3 and Proposition 4.4

Proof of Lemma 4.3. Since the lab faces weights $(v_1, \theta v_2)$ in channel O, the calculation in the proof of Lemma 4.2 gives the difference between its certainty equivalents in channels I and O as $\ell + (1-\theta)D_{12}v_1v_2 + \frac{1}{2}(1-\theta^2)(D_{22} + 1/c_g - r\sigma_2^2)v_2^2 = \Phi_\theta(v_2)$. The factored form follows from $1 - \theta^2 = (1-\theta)(1+\theta)$. Since $1 - \theta > 0$, $\Phi_\theta \geq 0$ is equivalent to $\ell_\theta - sv_2 + \frac{1}{2}K_\theta v_2^2 \geq 0$. The form of the set follows from the proof of Proposition 4.3(ii), which used only $s > 0$.

We next show that the set of weights inducing exit shrinks as θ increases. For $K \leq 0$, this follows from $\partial\Phi_\theta(v_2)/\partial\theta = sv_2 + \theta|K|v_2^2 \geq 0$. Let $K > 0$ and $\theta < \theta'$, and take a point v_2 that induces exit under θ', so $\Phi_{\theta'}(v_2) < 0$. Since $\Phi_{\theta'}(v_2)/(1-\theta') = \ell_{\theta'} - sv_2 + \frac{1}{2}(1+\theta')Kv_2^2 < 0$ and $\ell_{\theta'} \geq 0$, we have $v_2 < 2s/\{(1+\theta')K\} < s/(\theta' K)$. At this point, $\partial\Phi_{\theta''}(v_2)/\partial\theta'' = v_2(s - \theta'' K v_2) > 0$ for $\theta'' \in [\theta, \theta']$. Hence $\Phi_\theta(v_2) < \Phi_{\theta'}(v_2) < 0$, so v_2 also induces exit under θ. Finally, if $K \leq 0$ and $\ell > 0$, then $\partial\Phi_\theta(v_2)/\partial\theta > 0$ for $v_2 > 0$. Since v_θ^- is the unique positive root of Φ_θ, it is strictly increasing in θ. For $K > 0$, monotonicity of v_θ^- follows from the shrinking set of weights inducing exit. □

Proof of Proposition 4.4 For an announced weight $v_2 \in [0, \bar{W}]$, the lab chooses channel I if $\Phi_\theta(v_2) > 0$, channel O if $\Phi_\theta(v_2) < 0$, and the community's preferred channel under Assumption 4.7(d) if $\Phi_\theta(v_2) = 0$. The attainable outcomes are therefore channel I with $v_2 \in \mathcal{I}_\theta := \mathcal{V}_\theta \cap [0, \bar{W}]$, and channel O with $v_2 \in \mathcal{O}_\theta := \{v_2 \in [0, \bar{W}] : \Phi_\theta(v_2) \leq 0\}$. Their respective surpluses are $S(v_1, v_2)$ and $S(v_1, \theta v_2) - \zeta$.

(i)(ii) Since $\bar{W} < v_\theta^+$, $\mathcal{I}_\theta = [0, \min\{v_\theta^-, \bar{W}\}]$, and, if $v_\theta^- \leq \bar{W}$, $\mathcal{O}_\theta = [v_\theta^-, \bar{W}]$. (a) If $v_2^* \leq \min\{v_\theta^-, \bar{W}\}$, the maximum surplus $S(v_1, v_2^*)$ is attained in channel I. (b) If $v_\theta^- \geq \bar{W}$ and $v_2^* > \bar{W}$, then $\mathcal{I}_\theta = [0, \bar{W}]$ and $\mathcal{O}_\theta$ is either empty or $\{\bar{W}\}$. Even if exit is induced, the effective weight is only $\theta\bar{W} < \bar{W} < v_2^*$, so $\bar{W}$ in channel I is optimal. (c) If $v_\theta^- < \bar{W}$ and $v_2^* > v_\theta^-$, the best weight in channel I is v_θ^-. The best effective weight in channel O is the point of $[\theta v_\theta^-, \theta\bar{W}]$ closest to v_2^*, which equals $\min\{v_2^*, \theta\bar{W}\}$ because $\theta v_\theta^- \leq v_\theta^- < v_2^*$. It is implemented by the announced weight $\min\{v_2^*/\theta, \bar{W}\}$ (or $\bar{W}$ if $\theta = 0$), which belongs to $\mathcal{O}_\theta$ because $v_2^*/\theta > v_\theta^-$. Comparing the two outcomes gives the condition in (ii). When $\zeta = 0$, $S(v_1, \cdot)$ is strictly increasing on $[0, v_2^*]$ and $v_\theta^- < v_2^*$, so the condition is equivalent to $\min\{v_2^*, \theta\bar{W}\} > v_\theta^-$, or $\theta\bar{W} > v_\theta^-$. The expression in (i) follows because $v_2^* \leq R_\theta$ in (a), $R_\theta = \bar{W} < v_2^*$ in (b), and $R_\theta = \max\{v_\theta^-, \theta\bar{W}\}$ in (c).

(iii) Since the effective weight satisfies $\theta v_2 \le v_2 \le \bar{W}$ and $\zeta \ge 0$, $\mathcal{S}_\theta^* \le \max_{0\le x\le \bar{W}} S(v_1, x) = S(v_1, \min\{v_2^*, \bar{W}\})$. When $\bar{W} < v_\theta^+$, the proof of (i)(ii) gives

$$\mathcal{S}_\theta^* = \max\Big\{S(v_1, \min\{v_2^*, v_\theta^-, \bar{W}\}),\ X_\theta\Big\},$$
$$X_\theta \coloneqq \begin{cases} S(v_1, \min\{v_2^*, \theta\bar{W}\}) - \zeta & (v_\theta^- < \bar{W}) \\ -\infty & (v_\theta^- \ge \bar{W}). \end{cases}$$

This expression also covers case (a), where $X_\theta \le S(v_1, v_2^*)$. The first term is nondecreasing because v_θ^- is nondecreasing in θ. The term X_θ is nondecreasing wherever $v_\theta^- < \bar{W}$; once $v_\theta^- \ge \bar{W}$, the first term equals the upper bound $S(v_1, \min\{v_2^*, \bar{W}\})$. Thus $\mathcal{S}_\theta^*$ is nondecreasing. If $K \le 0$, $v_\theta^+ = \infty > \bar{W}$ for every θ. If $K > 0$, the range of θ satisfying $\bar{W} < v_\theta^+$ consists of a possibly empty interval containing zero, together with the portion $\theta > \theta_c$, because v_θ^+ is nonincreasing for $\theta \le \theta_c$ (Appendix 4.B.2). In the latter portion, $\mathcal{V}_\theta = [0, \infty)$ and the upper bound is attained. Thus $\mathcal{S}_\theta^*$ is nondecreasing throughout the stated range.

(iv) As $v_1 \to \infty$, $s \to \infty$ and $v_2^* \to \infty$. If $K \le 0$, $v_\theta^- \le \ell_\theta/s \to 0$. If $K > 0$, then $2K_\theta\ell_\theta < s^2$ for sufficiently large s, and $v_\theta^- = 2\ell_\theta/(s + \sqrt{s^2 - 2K_\theta\ell_\theta}) \le 2\ell_\theta/s \to 0$, while $v_\theta^+ \ge s/K_\theta \to \infty$ (Appendix 4.B.2). Thus, for sufficiently large v_1, $\bar{W} < v_\theta^+$ and $v_\theta^- < \bar{W} < v_2^*$, so the best weight in channel I is v_θ^-. Using $\Phi_\theta(v_\theta^-) = 0$, or $s\, v_\theta^- = \ell_\theta + \frac{1}{2}K_\theta(v_\theta^-)^2$, its gain is

$$\frac{N}{\Delta}v_\theta^- - \tfrac{1}{2}H_2(v_\theta^-)^2 = \ell_\theta + x_2^{\mathrm{FB}}v_\theta^- + \tfrac{1}{2}(K_\theta - H_2)(v_\theta^-)^2 \to \ell_\theta.$$

If $\theta > 0$, then also $v_\theta^- < \theta\bar{W}$ for sufficiently large v_1, and (ii) applies. When exit is induced, the announced weight is $\bar{W}$ and the effective weight is $\theta\bar{W}$. The expression for the gain follows from $S(v_1, x) - S(v_1, 0) = (N/\Delta)x - \frac{1}{2}H_2x^2$. Since $N/\Delta = x_2^{\mathrm{FB}} + s \to \infty$, this gain diverges and, for sufficiently large v_1, exceeds the gain from retaining the lab in channel I. Hence the lab chooses channel O. By contrast, $G = (N/\Delta)^2/(2H_2)$, so the ratio to G tends to zero. If $\theta = 0$, inducing exit leaves the effective weight at zero and surplus no greater than $S(v_1, 0) - \zeta$. The lab therefore remains in channel I, and the gain converges to $\ell_0 = \ell$. □

4.A.7 Proof of Proposition 4.5

(i) For fixed $u > 0$, $\Phi_\theta(u; j) = \ell - (1-\theta)(c_{12}v_1(j)/\Delta)u + \frac{1}{2}(1-\theta^2)Ku^2$ is strictly decreasing in $v_1(j)$, and $\Phi_\theta(u; j) < 0$ is equivalent to $v_1(j) > v_1^\dagger(u)$. Since the values $v_1(j)$ are distinct, at most one type satisfies $\Phi_\theta(u; j) = 0$. Under (T), if $\ell > 0$, $dv_1^\dagger/du = (\Delta/c_{12})(-\ell_\theta/u^2 + K_\theta/2)$ is nonpositive on $(0, \bar{W}]$, which proves the claim. If $\ell = 0$, (T) implies $K \le 0$, so $v_1^\dagger \le 0$ and all labs choose channel O.

(ii) Fix u and let $\tilde{v}_j \in \{u, \theta u\}$ denote the effective weight for type j. By (i), labs choosing channel I form the lower part of the distribution of v_1, so $\tilde{v}_j$ is nonincreasing in j, regardless of how the indifferent type is treated. Since v_j^* is increasing, Chebyshev's sum inequality gives $\sum_j \varpi_j v_j^* \tilde{v}_j \le (\mathrm{E}_\varpi v^*)\hat{v}$, where $\hat{v} \coloneqq \sum_j \varpi_j \tilde{v}_j$, and Jensen's inequality gives $\sum_j \varpi_j \tilde{v}_j^2 \ge \hat{v}^2$. Consequently,

$$\sum_j \varpi_j g_j(\tilde{v}_j) \le H_2\Big\{(\mathrm{E}_\varpi v^*)\hat{v} - \tfrac{1}{2}\hat{v}^2\Big\} \le \frac{H_2}{2}(\mathrm{E}_\varpi v^*)^2.$$

Accounting for the loss ζ of labs choosing channel O gives the claimed inequality. When $\zeta = 0$, equality requires equality in Jensen's inequality, so that $\tilde{v}_j$ is constant, together with $\hat{v} = \mathrm{E}_{\varpi} v^*$. Since $\mathrm{E}_{\varpi} v^* > 0$ implies $u > 0$, the effective weight $\tilde{v}_j$ can be constant only if all labs choose the same channel, yielding either (a) or (b). Conversely, equality holds in both (a) and (b).

If $\Phi_\theta(\mathrm{E}_{\varpi} v^*; J) \geq 0$, (i) implies that all labs can choose channel I at $u = \mathrm{E}_{\varpi} v^*$. Even if type J is indifferent, $\mathrm{E}_{\varpi} v^* \leq v_J^*$ implies $g_J(\mathrm{E}_{\varpi} v^*) \geq g_J(\theta \mathrm{E}_{\varpi} v^*) - \zeta$, so type J chooses channel I. Thus the upper bound is attained. If $\zeta > 0$, attaining the upper bound requires $\varpi_{\mathrm{O}}(u) = 0$ and $u = \mathrm{E}_{\varpi} v^*$, making the optimum unique. Finally, since $G_j = (H_2/2)(v_j^*)^2$, $\sum_j \varpi_j G_j - (H_2/2)(\mathrm{E}_{\varpi} v^*)^2 = (H_2/2)\mathrm{Var}_{\varpi}(v^*)$.

(iii) As $v_1(J) \to \infty$, $v_J^* \to \infty$, $v_\theta^-(J) \leq 2\ell_\theta / s_J \to 0$, and $v_\theta^+(J) \to \infty$ (the proof of Proposition 4.4(iv)). The participation boundaries of the other types do not depend on $v_1(J)$. Let B be an upper bound, over $[0, \bar{W}]$, on the absolute value of the weighted sum of the gains of the other types.

Suppose $\theta > 0$. For $u \leq v_\theta^-(J)$, where type J chooses channel I, $\Gamma_\theta(u) \leq B + \varpi_J H_2 v_J^* v_\theta^-(J)$. The right-hand side is bounded because v_J^*/s_J is bounded. In contrast, $\Gamma_\theta(\bar{W}) \geq -B + \varpi_J\{g_J(\theta \bar{W}) - \zeta\} \to \infty$. Hence, for sufficiently large $v_1(J)$, type J chooses channel O under the optimal u. Let p be the largest participation boundary of the other types that is below $\bar{W}$, or set $p = 0$ if there is none. For $u \in (v_\theta^-(J), p]$,

$$\begin{aligned}\Gamma_\theta(\bar{W}) - \Gamma_\theta(u) &\geq \varpi_J\{g_J(\theta \bar{W}) - g_J(\theta u)\} - 2B \\ &\geq \varpi_J H_2 \theta (\bar{W} - p)(v_J^* - \theta \bar{W}) - 2B,\end{aligned}$$

and the right-hand side diverges. For $u \in (\max\{p, v_\theta^-(J)\}, \bar{W})$, the channels chosen by the other types are constant, and $\Gamma_\theta'(u) \geq \varpi_J H_2 \theta (v_J^* - \theta \bar{W}) - H_2 \bar{W}$ becomes positive, so Γ_θ increases. Even if some type is indifferent at $\bar{W}$, the tie-breaking rule favorable to the community makes $\Gamma_\theta(\bar{W})$ at least as large as its left-hand limit. Thus $u^* = \bar{W}$ for sufficiently large $v_1(J)$. The channels of the other types follow from (i).

Suppose $\theta = 0$ and $\ell > 0$. For sufficiently large $v_1(J)$, type J chooses channel O for $u > v_0^-(J)$, and its contribution is the constant $-\varpi_J \zeta$, because $g_J(0) = 0$. Therefore, on $(v_0^-(J), \bar{W}]$, $\Gamma_0(u)$ equals the weighted sum of gains $\Gamma_{-J}(u)$ in the economy without type J, minus $\varpi_J \zeta$. An optimum u_{-J}^* of Γ_{-J} does not depend on $v_1(J)$. If $\ell > 0$, all labs choose channel I for sufficiently small $u > 0$, giving $\Gamma_{-J}(u) > 0 = \Gamma_{-J}(0)$, so $u_{-J}^* > 0$. Hence, for sufficiently large $v_1(J)$, $v_0^-(J) < u_{-J}^*$, and the maximum on $(v_0^-(J), \bar{W}]$ is $\Gamma_{-J}^* - \varpi_J \zeta$. On the other hand, for $u \leq v_0^-(J)$ and sufficiently large $v_1(J)$, all labs choose channel I because the other types' participation boundaries are positive and fixed. Since g_j increases on $[0, v_j^*]$, $\Gamma_0(u) \leq \Gamma_0(v_0^-(J)) = \sum_{j<J} \varpi_j g_j(v_0^-(J)) + \varpi_J g_J(v_0^-(J))$. On the right-hand side, the first term converges to zero and the second to $\varpi_J \ell$: in the proof of Proposition 4.4(iv), the gain from retaining the lab in channel I converges to $\ell_0 = \ell$. Comparing the two cases proves the claim.

(iv) Under $\theta' > \theta$, announce a norm u^* that is optimal under θ. By Lemma 4.3, the set of weights inducing exit shrinks with θ, so labs choosing channel I under θ can also choose channel I under θ', leaving their contributions unchanged. For a type j choosing channel O under θ, the assumption gives $\theta u^* < \theta' u^* < u^* \leq v_j^*$, and g_j increases on $[0, v_j^*]$. Its contribution therefore does not fall, whether it remains in channel O under θ' (because $g_j(\theta' u^*) \geq g_j(\theta u^*)$) or moves to channel I (because $g_j(u^*) \geq g_j(\theta u^*)$). Thus $\Gamma_{\theta'}^* \geq \Gamma_{\theta'}(u^*) \geq \Gamma_\theta^*$. If $u^* = 0$, all effective weights remain zero, so the contributions

are unchanged; the same comparison then uses weak inequalities. The example in Appendix 4.B.3 shows that monotonicity need not hold in general. □

Appendix 4.B Supplementary results for Section 4

This appendix provides additional cases, general solutions, and results on dependence on assumptions that were not included in Propositions 4.3–4.5, together with their proofs. Remarks on Propositions 4.1 and 4.2 follow their respective proofs in Appendix 4.A.

4.B.1 Further results on the participation constraint

Proposition 4.B.1 (Details of the participation constraint). *Under the assumptions of Proposition 4.3, the following statements hold.*

(i) (Optimal norm when $K > 0$.) If $K > 0$ and $\ell \geq s^2/(2K)$, then $\mathcal{V} = [0, \infty)$ and $v_2^{\mathrm{P}} = v_2^*$. If $K > 0$ and $\ell < s^2/(2K)$, then $\mathcal{V} = [0, v_-] \cup [v_+, \infty)$, where $v_\pm = (s \pm \sqrt{s^2 - 2K\ell})/K$. Norms of intermediate strength push the lab outside the norm, with exit interval (v_-, v_+). When $v_2^* \in (v_-, v_+)$, v_2^{P} equals v_+ if $v_2^* \geq s/K$ (strengthening the norm), and v_- if $v_2^* < s/K$ (weakening it).

(ii) (Exit threshold.) If $v_1^\dagger(v_2)$ in Proposition 4.3(iii) is nonpositive, the lab chooses channel O for every $v_1 > 0$. If $\ell > 0$ and $K > 0$, $v_1^\dagger$ is U-shaped in v_2, attaining its minimum at $v_2 = \sqrt{2\ell/K}$. Strengthening a common norm beyond this point instead reduces exit.

(iii) (Comparative statics.) The set $\mathcal{V}$ increases, in the sense of set inclusion, with ℓ and $1/c_g$, and decreases with v_1 and $r\sigma_2^2$.

(iv) (Monotonicity of the shortfall.) The shortfall $\frac{1}{2}H_2 \operatorname{dist}(v_2^*, \mathcal{V})^2$ is nondecreasing in v_1.

Proof. (i) The form of the set was established in the proof of Proposition 4.3(ii). When $v_2^* \in (v_-, v_+)$, the closest point is whichever of v_- and v_+ is closer to v_2^*. Their midpoint is s/K; if they are equally distant, choose the larger point.

(ii) If $v_1^\dagger \leq 0$, then $\Phi(v_2) = \ell + \frac{1}{2}Kv_2^2 - sv_2 < 0$ for every $s > 0$. When $\ell > 0$ and $K > 0$, $\ell/v_2 + (K/2)v_2$ is strictly convex for $v_2 > 0$ and attains its minimum at $v_2 = \sqrt{2\ell/K}$.

(iii) This follows from $\partial\Phi/\partial\ell = 1$, $\partial\Phi/\partial(1/c_g) = \frac{1}{2}v_2^2 \geq 0$, $\partial\Phi/\partial(r\sigma_2^2) = -\frac{1}{2}v_2^2 \leq 0$, and $\partial\Phi/\partial v_1 = -(c_{12}/\Delta)v_2 \leq 0$.

(iv) Since H_2 does not depend on v_1, it suffices to show that $\operatorname{dist}(v_2^*, \mathcal{V})$ is nondecreasing in s. If $K \leq 0$, then $\operatorname{dist} = (v_2^* - \bar{v})^+$, where v_2^* is increasing and $\bar{v}$ is nonincreasing. For $K > 0$, $\operatorname{dist} = 0$ as long as v_2^* lies outside the exit interval. Inside the interval, $\operatorname{dist} = \min\{v_2^* - v_-,\ v_+ - v_2^*\}$, and

$$\frac{dv_2^*}{ds} = \frac{\Delta}{c_1 + \Delta\psi_2} \leq \frac{1}{K}, \qquad \frac{dv_\pm}{ds} = \frac{1}{K}\Big(1 \pm \frac{s}{\sqrt{s^2 - 2K\ell}}\Big).$$

Here $c_1 + \Delta\psi_2 - \Delta K = 2\Delta r\sigma_2^2 \geq 0$, and $dv_+/ds \geq 2/K$, while $dv_-/ds \leq 0$. Thus both $v_2^* - v_-$ and $v_+ - v_2^*$ increase, and once v_2^* enters the interval, it remains there. Hence dist is nondecreasing. □

Remark (Interpreting Φ). By the envelope theorem, $\Phi'(0) = x_2^0$. To first order, the lab's expected recognition under a weak norm equals its support for understanding without a norm, x_2^0, multiplied by v_2; the exact expression is $v_2\{x_2^0 + (c_1/\Delta + 1/c_g)v_2\}$. If $c_{12} > 0$ and the market reward is positive, $x_2^0 < 0$, and a stronger market reward implies stronger criticism under the norm.

Remark (Assumptions and scope). (1) Proposition 4.B.1(i)(ii), Proposition 4.3(iii), and the interpretation in terms of criticism depend on placing the reference point for recognition at the baseline level (Assumption 4.6). If the reference point is relaxed by $m > 0$, so that recognition is $v_2(y_2 + m)$, Φ is obtained by replacing s with $s - m$, and the participation constraint is relaxed. A more lenient reference point is similar to granting a fixed amount of recognition. However, the m required to prevent exit varies across labs with x_2^0.

(2) For $K > 0$, the branch that retains the lab by strengthening the norm (v_+) assumes that the recognition weight is uncapped and costless to the community. Proposition 4.4 analyzes a capped weight.

(3) The second part of Proposition 4.3(iv) does not hold when $r\sigma_2^2 = 0$. In that case, $v_2^* - s/K = x_2^{\mathrm{FB}}/K$, so if $x_2^{\mathrm{FB}} > 0$, then $v_2^{\mathrm{P}} = v_+ \to \infty$ for sufficiently large v_1, and the gain $\ell + x_2^{\mathrm{FB}} v_+$ also diverges.

(4) By Proposition 4.B.1(iii), making apparent compliance cheaper (reducing c_g) expands the participation set. If the participation constraint binds, this raises the enforceable weight and may therefore increase the gain. For example, in specification (a) of Appendix 4.C, where $K = 4/3$, let $\ell = 0.1$ and $v_1 = 2$. Reducing c_g from 2 to 1 lowers the unconstrained gain G from 1.524 to 1.255 but raises the gain under the participation constraint from 0.816 to 0.988. The effect of c_g in Proposition 4.2(ii) applies when the participation constraint does not bind.

(5) If a norm can be tailored to each lab, as in Proposition 4.3(i), exit does not occur. The exit in Proposition 4.3(iii) arises because the norm is common.

Remark (Social cost of delay and losses specific to channel O). The parameter ℓ is treated as a private value, like recognition, and is not included in social surplus. If delay also entails a social cost $d > 0$, surplus in channel I becomes $S(v_1, v_2) - d$. Proposition 4.3(iv) then implies that, if $r\sigma_2^2 > 0$ and $d > \ell$, sufficiently intense competition makes it preferable for the community to permit exit rather than retain the lab in channel I. By contrast, introducing a social loss specific to channel O, such as the spread of errors without peer review, does not change the conclusion of Proposition 4.3(i). The community can always choose a weight that retains the lab in channel I.

4.B.2 Further results on partial compliance

Participation boundaries when $K > 0$ For $K > 0$, the participation set $\mathcal{V}_\theta = [0, v_\theta^-] \cup [v_\theta^+, \infty)$ in Lemma 4.3 is given by

$$v_\theta^\pm = \frac{s \pm \sqrt{s^2 - 2K_\theta \ell_\theta}}{K_\theta} \qquad (\text{if } 2K_\theta\ell_\theta \le s^2),$$
$$v_\theta^- = v_\theta^+ = \infty, \quad \mathcal{V}_\theta = [0, \infty) \qquad (\text{if } 2K_\theta\ell_\theta > s^2).$$

For either sign of K, $\{v_2 \ge 0 : \Phi_\theta(v_2) \le 0\} = \{v_2 \ge 0 : v_\theta^- \le v_2 \le v_\theta^+\}$. If $K > 0$ and $\ell > 0$,

$$2K_\theta\ell_\theta = \frac{2K\ell(1+\theta)}{1-\theta} < s^2 \iff \theta < \theta_c := \frac{s^2 - 2K\ell}{s^2 + 2K\ell}.$$

Thus weights inducing exit exist if and only if $\theta < \theta_c$. Since the set of such weights shrinks with θ, v_θ^+ is nonincreasing for $\theta \le \theta_c$. Also, for $\theta \ge \theta_c$, $\mathcal{V}_\theta = [0, \infty)$, with $v_\theta^- = v_\theta^+$ at $\theta = \theta_c$, so $\mathcal{S}_\theta^*$ equals the upper bound $S(v_1, \min\{v_2^*, \bar{W}\})$ in Proposition 4.4(iii).

An uncapped recognition weight When $\theta > 0$, the norm continues to affect a lab after it exits, which changes the community's problem itself. For example, if $K < 0$ and the recognition weight is uncapped, then, as long as the nominal weight v_2^*/θ lies outside the participation set, the community can announce it, induce exit, and impose the effective weight v_2^*. Thus Proposition 4.3(i) and (iv) do not extend unchanged to $\theta > 0$, which is why the main analysis imposes the cap $\bar{W}$. Partial compliance, such as disclosing a model name and some formalization while bypassing journals, can be interpreted as a situation with $0 < \theta < 1$.

Proposition 4.B.2 (General solution under partial compliance). *Under the assumptions of Proposition 4.4, define* $\mathcal{I}_\theta := \mathcal{V}_\theta \cap [0, \bar{W}]$ *and* $\mathcal{O}_\theta := \{v_2 \in [0, \bar{W}] : \Phi_\theta(v_2) \le 0\}$ *as in the proof of Proposition 4.4 in Appendix 4.A.*

(i) (General solution.) Irrespective of the ordering of the cap and the upper participation boundary,

$$\mathcal{S}_\theta^* = S(v_1, v_2^*) - \min\Big\{\tfrac{1}{2}H_2 \operatorname{dist}(v_2^*, \mathcal{I}_\theta)^2, \quad \tfrac{1}{2}H_2 \operatorname{dist}(v_2^*, \theta\mathcal{O}_\theta)^2 + \zeta\Big\},$$

where the second term is omitted if $\mathcal{O}_\theta = \emptyset$. If the first term is no greater than the second, the community announces a weight in $\mathcal{I}_\theta$ closest to v_2^*, and the lab chooses channel I. Otherwise, letting $\tilde{v}_2$ be a point in $\theta\mathcal{O}_\theta$ closest to v_2^*, it announces $\tilde{v}_2/\theta$, and the lab chooses channel O. If $\zeta = 0$, the set of attainable effective weights is

$$\mathcal{A}_\theta := \mathcal{I}_\theta \cup \theta\mathcal{O}_\theta = [0, R_\theta] \cup [v_\theta^+, \bar{W}],$$
$$R_\theta := \max\Big\{\min\{v_\theta^-, \bar{W}\},\ \theta \min\{v_\theta^+, \bar{W}\}\Big\},$$

and $\mathcal{S}_\theta^* = S(v_1, v_2^*) - \frac{1}{2}H_2 \operatorname{dist}(v_2^*, \mathcal{A}_\theta)^2$. If $\bar{W} < v_\theta^+$, R_θ agrees with the expression in Proposition 4.4(i).

(ii) (Range of θ in which a norm that is not fully followed is optimal.) Suppose $K \le 0$, $\ell > 0$, and $\zeta = 0$. The set of values of θ for which inducing exit is strictly preferable is an interval, possibly empty. Outside that interval, the lab remains in channel I.

Proof. (i) As established at the beginning of the proof of Proposition 4.4, the attainable outcomes are channel I with $v_2 \in \mathcal{I}_\theta$ and channel O with $v_2 \in \mathcal{O}_\theta$. Since $S(v_1, x) = S(v_1, v_2^*) - \frac{1}{2}H_2(x - v_2^*)^2$, $\mathcal{I}_\theta$ is compact and contains zero, and $\mathcal{O}_\theta$ is compact, the respective maxima are attained at points closest to v_2^*. Now let $\zeta = 0$. We have $\mathcal{I}_\theta = [0, \min\{v_\theta^-, \bar{W}\}] \cup [v_\theta^+, \bar{W}]$, where the second interval is empty if $v_\theta^+ > \bar{W}$. If $v_\theta^- \le \bar{W}$, $\mathcal{O}_\theta = [v_\theta^-, \min\{v_\theta^+, \bar{W}\}]$; if $v_\theta^- > \bar{W}$, $\mathcal{O}_\theta = \emptyset$. In the former case, $\theta v_\theta^- \le v_\theta^-$ implies $[0, v_\theta^-] \cup \theta\mathcal{O}_\theta = [0, R_\theta]$. In the latter case, $\mathcal{A}_\theta = [0, \bar{W}]$, and $\theta \min\{v_\theta^+, \bar{W}\} < \bar{W}$ gives $R_\theta = \bar{W}$.

(ii) By the proof of Proposition 4.4(i)(ii), inducing exit is strictly preferable if and only if $v_\theta^- < v_2^*$ and $\theta\bar{W} > v_\theta^-$. When $v_2^* \le \min\{v_\theta^-, \bar{W}\}$, exit may also implement the best weight, but cannot be strictly preferable. If $K \le 0$, Φ_θ is strictly decreasing for

$v_2 \geq 0$, so the latter condition is equivalent to $\Phi_\theta(\theta\bar{W}) < 0$. The formula for v_θ^- implies continuity in θ, and Lemma 4.3 implies strict monotonicity. Hence the set of θ satisfying the former condition is either empty or an interval containing zero. The latter condition can be written as $\chi(\theta) > \ell$, where

$$\chi(\theta) \coloneqq \ell - \Phi_\theta(\theta\bar{W}) = \bar{W}\Big\{s\,\theta(1-\theta) + \tfrac{1}{2}|K|\bar{W}\,\theta^2(1-\theta^2)\Big\}.$$

The derivative $\chi'(\theta)/\bar{W} = s(1-2\theta) + |K|\bar{W}\,\theta(1-2\theta^2)$ is positive for $\theta < 1/2$ and negative for $\theta \geq 1/\sqrt{2}$. On $[1/2, 1/\sqrt{2}]$, $\chi''(\theta)/\bar{W} = -2s + |K|\bar{W}(1-6\theta^2) < 0$, so χ' is strictly decreasing and χ' has exactly one zero. Thus χ first strictly increases and then strictly decreases on $[0,1)$, making $\{\theta : \chi(\theta) > \ell\}$ an open interval, possibly empty. The intersection of two intervals is an interval. □

Remark. (1) (Relation to Proposition 4.3.) If $\theta = 0$, then $\theta\mathcal{O}_\theta \subseteq \{0\} \subseteq \mathcal{I}_\theta$, so inducing exit can never be strictly preferable. Proposition 4.B.2(i) reduces to Proposition 4.3(i) with the participation set replaced by $\mathcal{V} \cap [0, \bar{W}]$. Because of the cap, the gain converges to ℓ as competition intensifies even if $r\sigma_2^2 = 0$ (Proposition 4.4(iv)); the exception in item (3) of the remark "Assumptions and scope" in Appendix 4.B.1 does not arise.

(2) (Retaining the lab in channel I.) Under $\bar{W} < v_\theta^+$, if the other conditions of Proposition 4.4(ii) are not satisfied, it is optimal to retain the lab in channel I: at weight v_2^* if $v_2^* \leq \min\{v_\theta^-, \bar{W}\}$, and at weight $\bar{W}$ if $v_\theta^- \geq \bar{W}$ and $v_2^* > \bar{W}$. This holds regardless of ζ; see cases (a) and (b) in the proof of Proposition 4.4 in Appendix 4.A.

(3) (A cap above the upper participation boundary.) If $K > 0$ and $\bar{W} \geq v_\theta^+$, the lab can also be retained in channel I by strengthening the norm to at least v_θ^+, corresponding to the second interval in $\mathcal{A}_\theta$. In this case, the monotonicity in Proposition 4.4(iii) need not hold. The upper bound on the effective weight following exit, θv_θ^+, may fall as θ approaches θ_c, because v_θ^+ falls rapidly. For example, take specification (a) in Appendix 4.C, with $K = 4/3$, and set $b_2 = 0.55$, $v_1 = 1.75$, $\ell = 0.1$, $\bar{W} = 3$, and $\zeta = 0$. Increasing θ from 0.56 to 0.66 raises the surplus shortfall $S(v_1, v_2^*) - \mathcal{S}_\theta^*$ from 0.6% to 4.2% of G. On the normalized parameter grid examined, declines of this kind did not exceed approximately 3.6% of G.

4.B.3 Further results on a common norm

Proposition 4.B.3 (General solution for a common norm). *Under the assumptions of Proposition 4.5, arrange 0, $\bar{W}$, and all values $v_\theta^\pm(j)$ in $(0, \bar{W})$ in increasing order. On each open interval between adjacent points, the set $\mathcal{C}$ of labs choosing channel I is constant, and Γ_θ equals*

$$\Gamma_{\mathcal{C}}(u) \coloneqq \sum_{j \in \mathcal{C}} \varpi_j g_j(u) + \sum_{j \notin \mathcal{C}} \varpi_j \{g_j(\theta u) - \zeta\}.$$

The function $\Gamma_{\mathcal{C}}$ is a concave quadratic, with maximizer

$$\bar{u}_{\mathcal{C}} = \frac{\sum_{j \in \mathcal{C}} \varpi_j\, v_j^* + \sum_{j \notin \mathcal{C}} \theta^2 \varpi_j\, (v_j^*/\theta)}{\sum_{j \in \mathcal{C}} \varpi_j + \sum_{j \notin \mathcal{C}} \theta^2 \varpi_j}.$$

At $\theta = 0$, each outside-channel term in the numerator is interpreted by continuous extension: $\theta^2\varpi_j(v_j^*/\theta) = \theta\varpi_j v_j^*$, which is zero. Thus, for $\theta > 0$, $\bar{u}_{\mathcal{C}}$ is a weighted average of v_j^* for labs choosing channel I, with weights ϖ_j, and the nominal targets v_j^*/θ for labs choosing channel O, with weights $\theta^2\varpi_j$. If $\theta = 0$ and $\mathcal{C} = \emptyset$, $\Gamma_{\mathcal{C}}$ is constant, so this ratio is not needed. If $\theta > 0$, $\bar{u}_{\mathcal{C}}$ increases with v_j^* for labs choosing channel O; if $\theta = 0$, it is the average of v_j^* for labs choosing channel I. The optimal common norm maximizes Γ_θ over the points $\bar{u}_{\mathcal{C}}$ lying in interval interiors and all interval endpoints. At endpoints, indifferent labs' channels are selected in the community's favor.

Proof. On each open interval between adjacent points, no type is indifferent, so the sign of $\Phi_\theta(u; j)$, and hence the channel chosen by each lab, is constant. The first-order condition for $\Gamma_{\mathcal{C}}$, $\sum_{j\in\mathcal{C}} \varpi_j(v_j^* - u) + \theta\sum_{j\notin\mathcal{C}} \varpi_j(v_j^* - \theta u) = 0$, gives $\bar{u}_{\mathcal{C}}$. The second derivative is $-H_2(\sum_{j\in\mathcal{C}} \varpi_j + \theta^2\sum_{j\notin\mathcal{C}} \varpi_j) \leq 0$. On the closure of each interval, the maximum is attained at $\bar{u}_{\mathcal{C}}$ if that point lies in the interior, and otherwise at an endpoint. At an endpoint, indifferent labs' channels can be chosen in the community's favor, so the value of Γ_θ is at least as high as the values given by the expressions for either adjacent interval. Since there are finitely many endpoints, an optimum lies among the stated candidates. □

Jumps in the weighted sum of gains The function Γ_θ may change discontinuously at a point where a lab changes channels. When type j exits at norm u_0, the jump is $\varpi_j\{g_j(\theta u_0) - g_j(u_0) - \zeta\} = \varpi_j\{H_2(1-\theta)u_0[(1+\theta)u_0/2 - v_j^*] - \zeta\}$, which is downward if $u_0 < 2v_j^*/(1+\theta)$. If the norm is too strong, with $u_0 > 2v_j^*/(1+\theta)$, the jump is upward when ζ is sufficiently small.

Remark. (1) (Relation to Proposition 4.4.) For $J = 1$, Proposition 4.5 and Proposition 4.B.3 reduce to Proposition 4.4 and Proposition 4.B.2(i). The upper bound in Proposition 4.5(ii) corresponds to the gain G in Proposition 4.2 when $J = 1$.

(2) (An example in which increasing θ reduces surplus.) Let $\ell = 0.3$ and $\zeta = 0$, and consider the same three lab types as in the numerical example of a common norm in Appendix 4.C, with $v_1(3) = 3$. For $0.469 \leq \theta \leq 0.548$, the allocation in which all labs exit under the announced weight $\mathrm{E}_\varpi v^*/\theta$ (Proposition 4.5(ii)(b)) attains the pooling bound of 0.477. As θ increases, the participation boundary $v_\theta^-(1)$ of type 1, which faces the weakest competition, rises above $\mathrm{E}_\varpi v^*/\theta$. Inducing type 1 to exit then requires strengthening the norm to that boundary, making the effective weight too large. Surplus falls to 0.451 at $\theta \approx 0.619$, where the optimum switches to an allocation that retains all labs in channel I. The upper bound is attained again for $\theta \geq 0.718$, where $v_\theta^-(3) \geq \mathrm{E}_\varpi v^*$. Thus increasing θ makes it harder to allow labs for which the norm is too strong to escape its full force. The value Γ_θ^* can also fall discontinuously with θ. This occurs when the allocation in which all labs exit ceases to be feasible, for example because type 1's participation boundary exceeds the cap $\bar{W}$ when $K \leq 0$. Among the examples examined, some declines exceeded half of the unconstrained optimum $\sum_j \varpi_j G_j$. Declines of around 10% were also observed when $\zeta > 0$ or when the initial allocation retained some labs in channel I.

4.B.4 A lower bound on support for understanding

This subsection establishes the results in Section 4.5.4. Let the baseline level be $L > 0$ and restrict the lab's choice to $x_2 \geq -L$, so that the total level is nonnegative. Write social surplus under effective weight $\tilde{v}_2$ as $S_L(v_1, \tilde{v}_2)$, with an additional deduction of ζ in channel O.

Proposition 4.B.4 (A lower bound on support for understanding). *Under Assumptions 4.1–4.5, suppose $c_{12} > 0$ and $v_1 > 0$, and restrict the lab's choice to $x_2 \geq -L$.*

(i) Under effective weight $\tilde{v}_2 \geq 0$, the lab's optimal actions are $g^* = \tilde{v}_2/c_g$, $x_2^* = \max\{-L,\ (c_1\tilde{v}_2 - c_{12}v_1)/\Delta\}$, and $x_1^* = (v_1 - c_{12}x_2^*)/c_1$.

(ii) Under Assumption 4.6 as modified by Assumption 4.7, suppose $v_1 \geq v_1^L \coloneqq (c_1\bar{W} + \Delta L)/c_{12}$. Then, in either channel, $x_2^* = -L$ for every $v_2 \in [0, \bar{W}]$, and x^* does not depend on v_2. Social surplus is $S_L(v_1, \tilde{v}_2) = S_L(v_1, 0) - \frac{1}{2}\psi_2\tilde{v}_2^2$, and the maximum surplus attainable by the community is $S_L(v_1, 0)$. Thus the gain from a norm is zero regardless of $\theta \in [0, 1)$ and $\zeta \geq 0$.

(iii) Under Assumption 4.6, with $\theta = 0$ and no cap on the recognition weight, suppose $K \leq 0$. Define $a \coloneqq r\sigma_2^2 - 1/c_g$, where $K \leq 0$ implies $a \geq c_1/\Delta > 0$, and

$$\bar{v}_L \coloneqq \frac{2\ell}{L + \sqrt{L^2 + 2a\ell}}.$$

If $v_1 > (c_1\bar{v}_L + \Delta L)/c_{12}$, the participation set is $[0, \bar{v}_L]$, throughout which $x_2^* = -L$. Consequently, the gain from a norm is zero.

(iv) Under Assumption 4.8, for any type j with $v_1(j) \geq v_1^L$, $S_j \leq S_L(v_1(j), 0)$ under every common norm $u \in [0, \bar{W}]$. Thus the gain for such a type cannot be positive under any common norm.

Proof. (i) The choice of g is separable from the other choices, giving $g^* = \tilde{v}_2/c_g$. The objective in x, $v_1x_1 + \tilde{v}_2x_2 - \frac{1}{2}x^\top Cx$, is strictly concave, and the optimal x_1 for fixed x_2 is $(v_1 - c_{12}x_2)/c_1$. After substitution, the objective is a concave quadratic in x_2 with unconstrained maximizer $(c_1\tilde{v}_2 - c_{12}v_1)/\Delta$. Its maximizer subject to $x_2 \geq -L$ is therefore the larger of this point and $-L$.

(ii) Since $\tilde{v}_2 \leq v_2 \leq \bar{W}$ and $v_1 \geq v_1^L$, $c_1\tilde{v}_2 - c_{12}v_1 \leq c_1\bar{W} - c_{12}v_1 \leq -\Delta L$. By (i), $x_2^* = -L$ and $x_1^* = (v_1 + c_{12}L)/c_1$, neither of which depends on $\tilde{v}_2$. The only components of social surplus that depend on $\tilde{v}_2$ are the cost of apparent compliance, $\frac{1}{2}c_g(g^*)^2 = \tilde{v}_2^2/(2c_g)$, and the risk premium, $\frac{r}{2}\sigma_2^2\tilde{v}_2^2$. Hence $S_L(v_1, \tilde{v}_2) = S_L(v_1, 0) - \frac{1}{2}\psi_2\tilde{v}_2^2 \leq S_L(v_1, 0)$. Channel O entails the further deduction $\zeta \geq 0$. On the other hand, announcing $v_2 = 0$ leads the lab to choose channel I because $\Phi_\theta(0) = \ell \geq 0$, with ties resolved by Assumption 4.7(d), and yields surplus $S_L(v_1, 0)$.

(iii) Let $\hat{v} \coloneqq (c_{12}v_1 - \Delta L)/c_1$. By assumption, $\hat{v} > \bar{v}_L \geq 0$. In channel O, where the weight is zero, $x_2^* = -L$. If $v_2 \leq \hat{v}$, then $x_2^* = -L$ in channel I as well, and x_1^* is also the same. The difference between the certainty equivalents in the two channels is therefore $\Phi_L(v_2) \coloneqq \ell - Lv_2 - \frac{a}{2}v_2^2$. The function Φ_L is strictly decreasing, and $\bar{v}_L$ is its nonnegative root, which is positive if $\ell > 0$. For $v_2 \geq \hat{v}$, the lower bound does not constrain channel I, so the difference equals $\Phi(v_2)$ from Lemma 4.2 plus the reduction

in the certainty equivalent caused by the binding lower bound in channel O, a constant independent of v_2. Since $K \leq 0$ implies $\Phi'(v_2) = -s + Kv_2 < 0$, the difference also decreases over this range. It is continuous at $\hat{v}$ and is therefore negative for all $v_2 > \bar{v}_L$. Thus the participation set is $[0, \bar{v}_L] \subset [0, \hat{v})$, on which x^* is independent of v_2. As in (ii), surplus in channel I is $S_L(v_1, 0) - \frac{1}{2}\psi_2 v_2^2$, and surplus in channel O is $S_L(v_1, 0)$, so the gain is zero.

(iv) Apply the argument in (ii) to type j. □

In the numerical example using the specification of Figure 2(a), with $\ell = 0.1$, $\bar{W} = 1$, and $\zeta = 0$, setting $L = 3$ gives $v_1^L = 6.5$. At $v_1 = 8$, the gain from a norm is zero for each of $\theta = 0$, 0.2, and 0.5. In the model without the lower bound, the respective gains are 0.124, 1.247, and 2.792.

Appendix 4.C Details of the numerical examples in Section 4

As in Section 4.6, the baseline parameters are $c_1 = c_2 = 1$, $c_{12} = 0.5$ ($\Delta = 0.75$), $b_1 = 1$, $r = 1$, and $\sigma_1^2 = 0.1$. For Proposition 4.3 onward, set $b_2 = 1.5$ and $c_g = 2$ (so that $x_2^{\mathrm{FB}} = 4/3$), and consider two specifications of measurement noise. **Specification (a)** sets $\sigma_2^2 = 0.5$ ($\psi_2 = 1$, $K = 4/3 > 0$), whereas **specification (b)** sets $\sigma_2^2 = 2.5$ ($\psi_2 = 3$, $K = -2/3 < 0$). Figure 2 in the main text uses specification (b). The baseline levels are assumed to be large enough to keep total levels nonnegative over the range displayed in each figure (see Appendix 4.B.4 for the case in which the lower bound on support for understanding binds).

4.C.1 Propositions 4.1 and 4.2 and Corollary 4.1

Table 5 reports the main values. Regarding the identity $G = v_2^* \cdot N/(2\Delta)$ in Remark (2) of Appendix 4.A.3, this example has $N/\Delta = 2$, so that $G = v_2^*$; hence the two rows of the table coincide.

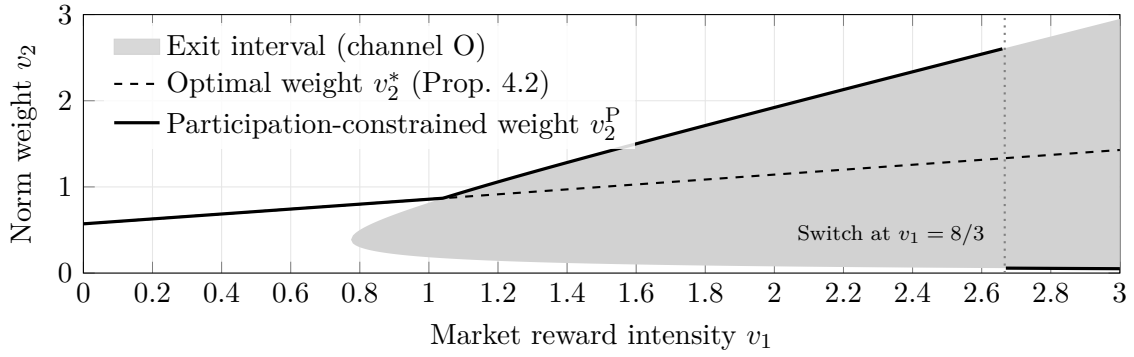
Table 5: Numerical examples for Propositions 4.1 and 4.2 and Corollary 4.1

Quantity	Value
Proposition 4.1 ($v_2 = 0$)	
$\rho = 0.5$: congruity κ_1, optimal reward v_1°, $S(v_1^\circ)$, $S(1)$	0.866, 0.698, 0.349, 0.283
$\rho = 1.5$: congruity κ_1, optimal reward v_1°, $S(v_1^\circ)$, $S(1)$	0.189, 0.233, 0.039, -0.383
$\rho = 2.5$: congruity κ_1, $S(1)$ ($S < 0$ for every $v_1 > 0$)	-0.115, -1.050
Effort (x_1^*, x_2^*) at $v_1 = 1$	(1.333, -0.667)
Proposition 4.2 ($\rho = 1.5$, $v_1 = 1$, $N = 1.5$)	
Optimal weight v_2^* ($\psi_2 = 0, 1, 2, 5$)	1.500, 0.857, 0.600, 0.316
Gain G ($\psi_2 = 0, 1, 2, 5$)	1.500, 0.857, 0.600, 0.316
Effort ($= x^{\mathrm{FB}}$) and remaining loss at $\psi_2 = 0$	(0.333, 1.333), 0.050 (risk premium only)
Shortfall under the cap $\bar{w}_2 = 0.6$ ($\psi_2 = 0, 1$)	0.540, 0.077
$v_1 = 1.5$, $\psi_2 = 0$: v_2^* and remaining loss	1.750, $0.125 + 0.1125$
Corollary 4.1 ($c_2 = c_3 = 1$, $c_{23} = 0.6$, $b_2 = 1$, $\psi_2 = 0$)	
Threshold above which the norm is harmful, $b_3/b_2 = c_3/c_{23}$	1.667
$b_3 = 2.0$: optimal norm weight	0 (every $v_2 > 0$ is harmful)
$b_3 = 1.5$: v_2^* and change in surplus under the naive weight v_2^{n}	0.100, -0.625
Threshold above which the naive weight is harmful, $b_3/b_2 > c_3/(2c_{23})$	0.833

4.C.2 Proposition 4.3: the participation constraint

Set the legitimacy value to $\ell = 0.1$ in specifications (a) and (b). For each market reward v_1, Figure 4 depicts the region in which the lab exits the norm-governed channel, the optimal weight v_2^* (Proposition 4.2), and the optimal weight subject to participation, v_2^{P}. Table 6 reports the main values. The upper bound $\ell + (x_2^{\mathrm{FB}})^2/(4r\sigma_2^2)$ in Proposition 4.3(iv) is 0.989 in specification (a) and 0.278 in specification (b).

(a) Specification (a), $K > 0$: intermediate norms induce exit



(b) Specification (b), $K < 0$: enforcement capacity as a cap $\bar{v}$

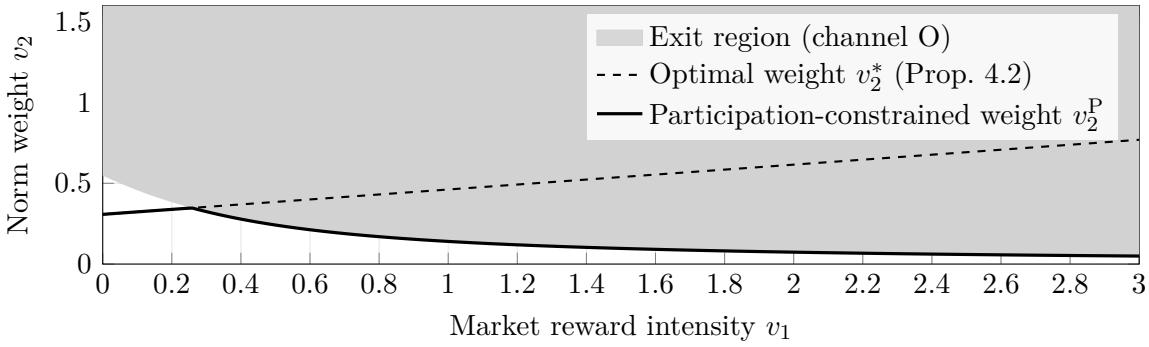


Figure 4: Enforcement capacity as a participation constraint (Proposition 4.3, $\ell = 0.1$). (a) Beyond $v_1 \approx 1.04$, the optimal weight v_2^* enters the exit interval. The community strengthens the norm (to the upper boundary v_+) to retain the lab in channel I, but switches to the lower boundary v_- when $v_1 > 8/3$ (dotted line); the norm then weakens toward zero (Proposition 4.B.1(i)). (b) Enforcement capacity takes the form of a cap $\bar{v}$, which becomes binding at $v_1 \approx 0.26$. The cap $\bar{v}$ decreases with v_1.

Table 6: Numerical examples of the participation constraint (Proposition 4.3, $\ell = 0.1$)

v_1	v_2^*	Participation set $\mathcal{V}$	v_2^{P}	Shortfall	Gain retained
Specification (a): $K = 4/3$ ($\sigma_2^2 = 0.5$)					
1.0	0.857	$[0, 0.184] \cup [0.816, \infty)$	0.857	0	100.0%
1.5	1.000	$[0, 0.108] \cup [1.392, \infty)$	1.392	0.180	84.6%
2.0	1.143	$[0, 0.078] \cup [1.922, \infty)$	1.922	0.708	53.5%
3.0	1.429	$[0, 0.051] \cup [2.949, \infty)$	0.051	2.214	7.0%
Specification (b): $K = -2/3$ ($\sigma_2^2 = 2.5$)					
0.5	0.385	$[0, 0.242]$	0.242	0.044	86.2%
1.0	0.462	$[0, 0.140]$	0.140	0.224	51.5%
2.0	0.615	$[0, 0.074]$	0.074	0.636	22.5%
3.0	0.769	$[0, 0.050]$	0.050	1.122	12.5%

The shortfall is $\frac{1}{2}H_2 \operatorname{dist}(v_2^*, \mathcal{V})^2$, and the fraction of the gain retained is $\{S(v_1, v_2^{\mathrm{P}}) - S(v_1, 0)\}/G$. In specification (a), the gain at $v_1 = 3$ is 0.167, consistent with the expression in Proposition 4.3(iv): $\ell + x_2^{\mathrm{FB}} v_2^{\mathrm{P}} - r\sigma_2^2 (v_2^{\mathrm{P}})^2 = 0.1 + 1.333 \times 0.051 - 0.5 \times 0.051^2$. Under common norms $v_2 = 0.3, 0.6, 0.9$, labs in specification (a) exit when $v_1 > 0.80, 0.85, 1.07$, respectively (Proposition 4.3(iii)). At these values, stronger common norms lead to less exit (Proposition 4.B.1(ii)).

4.C.3 Proposition 4.4: partial compliance

In specification (b), set $v_1 = 2$, $\ell = 0.1$, $\bar{W} = 1$, and $\zeta = 0$. Then $v_2^* = 0.615$ and $G = 0.821$. For each θ, Figure 5 depicts the participation boundary v_θ^-, the effective weight $\theta\bar{W}$ when exit is induced, the optimal effective weight $\tilde{v}_2$, and the announced weight. Table 7 reports the main values.

Inducing the lab to exit is strictly preferable when $0.080 < \theta < 0.906$ (Proposition 4.B.2(ii)). When $\theta < 0.080$, the norm is weakened to the participation boundary v_θ^- to retain the lab in channel I. When $0.080 < \theta < 0.615$, the community announces the cap $\bar{W} = 1$, induces exit, and imposes the effective weight θ. Once θ exceeds $v_2^*/\bar{W} = 0.615$, the optimal weight v_2^* can be imposed effectively while the lab remains outside (the announced value is v_2^*/θ). When $\theta > 0.906$, $v_\theta^- \geq v_2^*$, so the optimal weight can be imposed while retaining the lab in channel I. Surplus is nondecreasing in θ (Proposition 4.4(iii)). At $\theta = 0.3$, the gain is 0.605 when exit is induced and 0.253 when the lab is retained in channel I; hence, when $\zeta > 0$, inducing exit is preferable if $\zeta < 0.352$ (Proposition 4.4(ii)).

Figure 2(a) in the main text depicts the gain from the norm as a function of v_1 for $\theta = 0$, 0.2, 0.5, under the same specification. In specification (a) ($K = 4/3$), setting $v_1 = 2$ and $\bar{W} = 1$ makes inducing exit strictly preferable when $0.086 < \theta < 0.613$. For $\theta \geq 0.613$, $v_\theta^+ \leq \bar{W}$, so the norm can be strengthened to the cap $\bar{W}$ while retaining the lab in channel I (the gain reaches its capped maximum of 1.500; Proposition 4.B.2(i)).

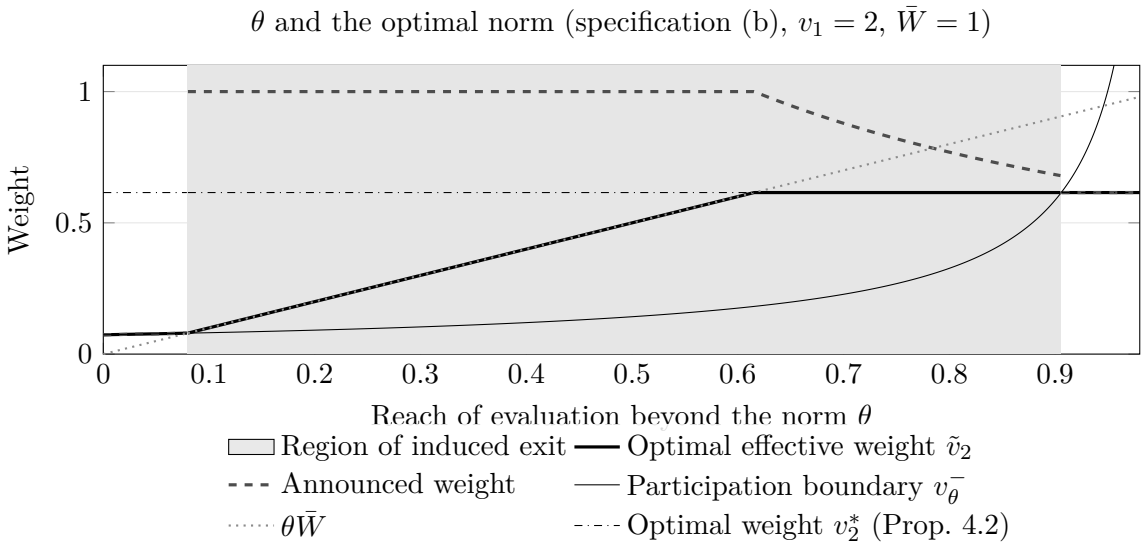


Figure 5: Partial compliance and the optimal norm (Proposition 4.4; specification (b), $v_1 = 2$, $\ell = 0.1$, $\bar{W} = 1$, $\zeta = 0$). In the shaded region, the community announces a strong norm at or below the cap to induce exit and imposes an effective weight through evaluation extending beyond the norm-governed channel. For $0.615 < \theta < 0.906$, the announced weight is v_2^*/θ.

Table 7: Numerical examples of partial compliance (Proposition 4.4; specification (b), $v_1 = 2$, $\ell = 0.1$, $\bar{W} = 1$, $\zeta = 0$)

θ	v_θ^-	$\theta\bar{W}$	Channel	Announced weight	$\tilde{v}_2$	Gain	Gain retained
0.00	0.074	0.000	I	0.074	0.074	0.185	22.5%
0.05	0.077	0.050	I	0.077	0.077	0.193	23.6%
0.10	0.082	0.100	O	1.000	0.100	0.245	29.9%
0.30	0.104	0.300	O	1.000	0.300	0.605	73.7%
0.50	0.142	0.500	O	1.000	0.500	0.792	96.5%
0.70	0.228	0.700	O	0.879	0.615	0.821	100.0%
0.90	0.587	0.900	O	0.684	0.615	0.821	100.0%
0.95	1.006	0.950	I	0.615	0.615	0.821	100.0%

The announced weight is the publicly stated norm weight, and $\tilde{v}_2$ is the effective weight. The gain is $\mathcal{S}_\theta^* - S(v_1, 0)$, and the fraction retained is its ratio to $G = 0.821$. In the rows for channel O, the lab exits rather than complying with the norm, and the community imposes the effective weight $\theta\times$(announced weight) through evaluation extending beyond the norm-governed channel.

4.C.4 Proposition 4.5: a common norm

In specification (b), set $\ell = 0.5$, $\bar{W} = 1$, and $\zeta = 0.05$. There are three types of labs: $v_1(1) = 0.5$ (weight 0.6), $v_1(2) = 1.5$ (weight 0.3), and $v_1(3)$ (weight 0.1). Type 3 represents the most intensely competing lab, and $v_1(3)$ is varied. The values of v_j^* for types 1 and 2 are 0.385 and 0.538, respectively.

Figure 6 depicts $\Gamma_\theta(u)$ at $v_1(3) = 8$ for $\theta = 0$ and $\theta = 0.5$. The function Γ_θ changes discontinuously at points where a lab switches channels. The size of each jump is given by the expression in Appendix 4.B.3. Most jumps are downward, but a jump is upward when the norm is excessively strong ($u \approx 0.823$ at $\theta = 0$). When $\theta = 0$, it is optimal to induce type 3 to exit and to impose 0.436, the weighted average of v_j^* for types 1 and 2 (Proposition 4.B.3). In this example, $\varpi_3(\ell + \zeta) = 0.055$ is smaller than $\Gamma_{-3}^* = 0.371$, corresponding to the first case in Proposition 4.5(iii). Type 3 also exits when $\theta = 0.5$, but its apparent target raises the norm to 0.507.

Figure 2(b) in the main text depicts the optimal common norm as a function of $v_1(3)$. When $\theta = 0$ and $v_1(3) < 2.97$, the norm is weakened to type 3's participation boundary to retain all labs in channel I. For $v_1(3) \geq 2.97$, type 3 is induced to exit, and the norm is tailored to types 1 and 2 (remaining constant at 0.436). When $\theta > 0$, the norm is likewise weakened to type 3's participation boundary while type 3 is retained in channel I ($v_1(3) < 3.34$ for $\theta = 0.2$, and $v_1(3) < 4.61$ for $\theta = 0.5$). After type 3 exits, more intense competition for that type raises the norm. When $\theta = 0.2$, the norm reaches the cap $\bar{W}$ at $v_1(3) \geq 20.8$, and all labs exit. When $\theta = 0.5$ and $v_1(3) \geq 32.0$, the norm exceeds type 2's participation boundary (0.732), reaching 0.818, so type 2 also exits. Table 8 reports the main values. When $\theta > 0$, the difference relative to lab-specific norms (Proposition 4.B.2(i)) increases as competition for type 3 intensifies. When $\theta = 0$, by contrast, the difference decreases slightly because the exited type 3 does not affect the level of the norm.

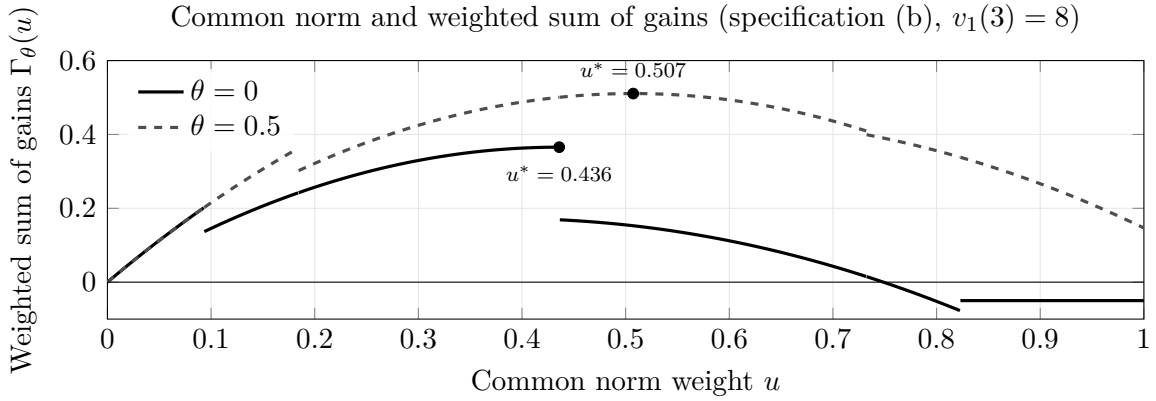


Figure 6: Common norm and weighted sum of gains (Proposition 4.5; specification (b), $\ell = 0.5$, $\bar{W} = 1$, $\zeta = 0.05$; weights are 0.6, 0.3, and 0.1 for types 1, 2, and 3, respectively). The function Γ_θ changes discontinuously where a lab switches channels (usually downward, but upward when the norm is excessively strong). The dots indicate optimal norms.

Table 8: Numerical examples of a common norm (Proposition 4.5; specification (b), $\ell = 0.5$, $\bar{W} = 1$, $\zeta = 0.05$)

$v_1(3)$	θ	u^*	Channels (types 1, 2, 3)	Γ^*_θ	Lab-specific norms	Pooling bound
3	0.0	0.436	I, I, O	0.366	0.442	0.477
3	0.2	0.295	I, I, I	0.411	0.460	0.477
3	0.5	0.449	I, I, I	0.476	0.488	0.477
8	0.0	0.436	I, I, O	0.366	0.434	0.646
8	0.2	0.468	I, I, O	0.424	0.500	0.646
8	0.5	0.507	I, I, O	0.511	0.655	0.646
30	0.0	0.436	I, I, O	0.366	0.427	1.696
30	0.2	1.000	O, O, O	0.630	0.793	1.696
30	0.5	0.690	I, I, O	0.950	1.388	1.696

Γ^*_θ is the weighted sum of gains under the common norm. "Lab-specific norms" reports the weighted sum of gains when the norm can be chosen separately for each lab (Proposition 4.B.2(i)). The pooling bound is $(H_2/2)(\mathrm{E}_\varpi v^*)^2$ in Proposition 4.5(ii). At $v_1(3) = 3$, $\theta > 0$ raises type 3's participation boundary, making it optimal to set the norm at that boundary and retain all labs in channel I.

Appendix 4.D Notation for Section 4

Table 9 summarizes the main notation used in Section 4.

Table 9: Notation for Section 4

Symbol	Meaning (first occurrence)
$x_1,\ x_2$	Effort devoted to solving and to support for understanding (deviations from baseline levels; Assumption 4.1)
g	Apparent compliance (activity that affects y_2 without contributing to understanding; Assumption 4.3)
$c_1,\ c_2,\ c_{12},\ \Delta$	Cost coefficients (efforts are substitutes when $c_{12} > 0$); $\Delta = c_1 c_2 - c_{12}^2$
$b_1,\ b_2,\ \rho$	Social-value coefficients; relative importance of support for understanding, $\rho = b_2/b_1$
$y_1,\ y_2,\ \sigma_1^2,\ \sigma_2^2$	Measures of problems solved and support for understanding, and their respective noise variances
$v_1,\ v_2$	Market reward intensity for solving (given); intensity of recognition awarded by the norm to support for understanding
$r,\ c_g,\ \psi_2$	Risk aversion; cost coefficient of apparent compliance; difficulty of verification, $\psi_2 = r\sigma_2^2 + 1/c_g$ (with $1/c_g = 0$ when $c_g = \infty$)
$S,\ S^{\mathrm{FB}},\ x^{\mathrm{FB}}$	Social surplus, first-best surplus, and first-best actions (Lemma 4.1)
$A,\ H_1,\ \kappa_1,\ v_1^{\circ}$	Marginal social value of the reward; curvature of surplus; congruity of the solving measure; socially optimal reward (Proposition 4.1)
$N,\ v_2^*,\ G,\ H_2$	$N = \Delta\{x_2^{\mathrm{FB}} - x_2^*(v_1, 0)\}$; optimal weight; gain from the norm; $H_2 = c_1/\Delta + \psi_2$ (Proposition 4.2)
$\bar{w}_2$	Cap on the weight the norm can impose (enforcement capacity; Proposition 4.2(iv); endogenized in Section 4.5)
ℓ	Legitimacy value of publication through the community channel (Assumption 4.6)
$\Phi,\ K,\ s,\ x_2^0$	Participation condition; its curvature $K = c_1/\Delta + 1/c_g - r\sigma_2^2$; $s = -x_2^0 = c_{12}v_1/\Delta$ (Lemma 4.2)
$\mathcal{V},\ \bar{v},\ v_{\pm},\ v_2^{\mathrm{P}}$	Participation set; its boundaries (for $K \le 0$ and $K > 0$, respectively); optimal weight subject to participation (Proposition 4.3)
$v_1^{\dagger}$	Market reward threshold for exit under a common norm (Proposition 4.3(iii), Proposition 4.5(i))
$\theta,\ \bar{W},\ \zeta$	Reach of evaluation beyond the norm-governed channel; cap on the announced recognition weight; social loss specific to channel O (Assumption 4.7)
$\Phi_\theta,\ \mathcal{V}_\theta,\ \ell_\theta,\ K_\theta,\ v_\theta^{\pm}$	Participation condition under partial compliance; participation set and its boundaries; $\ell_\theta = \ell/(1-\theta)$; $K_\theta = (1+\theta)K$ (Lemma 4.3)
θ_c	Threshold of θ above which no weight induces exit when $K > 0$ (Appendix 4.B.2)

(Table 9 continued)

Symbol	Meaning (first occurrence)
$\tilde{v}_2,\ R_\theta,\ \mathcal{S}^*_\theta$	Effective weight; effective enforcement capacity; maximum surplus attainable by the community (Proposition 4.4)
$\mathcal{I}_\theta,\ \mathcal{O}_\theta,\ \mathcal{A}_\theta$	Sets of weights that retain the lab in channel I, weights that induce exit, and attainable effective weights (proof of Proposition 4.4 in Appendix 4.A; Proposition 4.B.2(i))
$J,\ \varpi_j,\ v^*_j,\ u$	Number of lab types; type weight; optimal weight for type j; common norm (Assumption 4.8)
$\mathrm{E}_\varpi,\ \mathrm{Var}_\varpi$	Mean and variance under the weights ϖ_j (Section 4.5.3)
$g_j,\ \Gamma_\theta,\ \Gamma^*_\theta,\ \varpi_{\mathrm{O}}$	Gain for type j; weighted sum of gains and its maximum; weight of labs choosing channel O (Proposition 4.5)
Γ^*_{-J}	Maximum weighted sum of gains in the economy excluding type J, with weights left unchanged (Proposition 4.5(iii))
$x_3,\ c_3,\ c_{23},\ b_3,\ \Delta_{23}$	Support that is difficult to verify; its cost coefficients and social value; $\Delta_{23} = c_2c_3 - c_{23}^2$ (Corollary 4.1)
$M,\ v_2^{\mathrm{n}}$	Derivative of surplus at $v_2 = 0$; weight chosen without accounting for the value of x_3 (Corollary 4.1)
$L,\ S_L$	Baseline level defining the lower bound on support for understanding ($x_2 \geq -L$); social surplus under the lower bound (Proposition 4.B.4)
$v_1^L,\ \bar{v}_L,\ a$	Market reward level above which even a norm at the cap $\bar{W}$ cannot change support for understanding; participation boundary under the lower bound; $a = r\sigma_2^2 - 1/c_g$ (Proposition 4.B.4)

Appendix 5.A Proofs for Section 5

5.A.1 Lemma 5.1 and preliminaries

Proof of Lemma 5.1. Since $u''(a) = -(w + c_0) < 0$, u is strictly concave in a. The derivative $u'(a) = w(e - a) - \tau - c_0 a$ is decreasing, with $u'(0) = we - \tau$. If $we - \tau \leq 0$ (equivalently, $e \leq \bar{e}$), then $u'(a) \leq u'(0) \leq 0$ for $a \geq 0$, so $a^* = 0$. If $we - \tau > 0$, the solution to $u'(a) = 0$, namely $a = (we - \tau)/(w + c_0)$, is positive and, by strict concavity, is the unique maximizer. It can be written as $(we - \tau)/(w + c_0) = \omega(e - \bar{e})$. Finally, if $e > 0$, then $a^*(e) \leq \omega e < e$. □

Lemma 5.A.1 (Infinite Products). *Let $x_s \in [0, 1)$ for $s = 0, 1, 2, \ldots$, and define $P_t \coloneqq \prod_{s=0}^{t-1}(1 - x_s)$ ($P_0 = 1$). Since P_t is positive and nonincreasing, it has a limit $P_\infty \in [0, 1]$.*

(a) If $\sum_s x_s = \infty$, then $P_\infty = 0$.

(b) If $\sup_s x_s < 1$ and $\sum_s x_s < \infty$, then $P_\infty > 0$.

(c) If $\sum_s x_s < \infty$, then $\prod_{s \geq T}(1 - x_s) > 0$ for every T.

Proof. (a) The inequality $1 - x \leq e^{-x}$ gives $P_t \leq \exp(-\sum_{s<t} x_s) \to 0$. (b) Let $c \coloneqq \sup_s x_s < 1$. For $x \in [0, c]$, $-\log(1 - x) = \sum_{k \geq 1} x^k/k \leq x/(1 - x) \leq x/(1 - c)$. Hence $-\log P_t \leq \sum_s x_s/(1 - c) < \infty$, implying $P_\infty \geq \exp(-\sum_s x_s/(1 - c)) > 0$. (c) If $\sum_s x_s < \infty$, then $x_s \to 0$, so there exists S such that $x_s \leq 1/2$ for all $s \geq S$. The product $\prod_{s \geq T}(1 - x_s)$ is the product of finitely many positive numbers, $\prod_{T \leq s < \max\{T,S\}}(1 - x_s)$, multiplied by $\prod_{s \geq \max\{T,S\}}(1 - x_s)$, which is positive by part (b). □

5.A.2 Proof of Proposition 5.1

(i) Lemma 5.1 gives $a_t^* = \omega(e_t - \bar{e})^+$. If $e_t \leq \bar{e}$, then $a_t^* = 0$, and Assumption 5.3 gives $e_{t+1} = e_t$. If $e_t > \bar{e}$, then $e_{t+1} = e_t - \lambda_t \omega(e_t - \bar{e})$, so $e_{t+1} - \bar{e} = (1 - \lambda_t \omega)(e_t - \bar{e})$. Since $0 \leq \lambda_t \omega \leq \omega < 1$, we have $1 - \lambda_t \omega \in (0, 1]$, implying $e_{t+1} - \bar{e} \in (0, e_t - \bar{e}]$. Thus $e_{t+1} \leq e_t$ in either case. By induction, if $e_0 \leq \bar{e}$, then $e_t = e_0$ for every t; if $e_0 > \bar{e}$, then (5.3) holds for every t, with $a_t^* = \omega(e_t - \bar{e}) > 0$.

(ii) Since e_t is nonincreasing and bounded below, it has a limit. If $e_0 > \bar{e}$, (5.3) gives $e_\infty = \bar{e} + (e_0 - \bar{e})P_\infty$. Apply Lemma 5.A.1 to $x_s = \lambda_s \omega \in [0, \omega]$, where $\omega < 1$. Parts (a) and (b) imply $P_\infty = 0 \iff \sum_s \lambda_s \omega = \infty \iff \sum_s \lambda_s = \infty$. If $e_0 \leq \bar{e}$, then $e_\infty = e_0$.

(iii) This follows from $P_t \leq (1 - \underline{\lambda}\omega)^t$.

(iv) Since $e_t \geq 0$ is nonincreasing, V_t is also nonincreasing, and continuity gives $V_t \to \frac{1}{2}e_\infty^2$. Part (ii) implies $e_\infty \geq \min\{e_0, \bar{e}\}$ in all cases. If $\tau > 0$, then $\bar{e} > 0$, so $V_\infty > 0$. If $\sum_t \lambda_t = \infty$, then $e_\infty = \min\{e_0, \bar{e}\}$, and continuity of a^* gives $a_t^* \to a^*(e_\infty) = 0$.

(v) If $e \leq \bar{e}$, then $a^*(e) = 0$ and $e_{t+1} = e$. If $e > \bar{e}$ and $\lambda_t > 0$, then $e_{t+1} < e$. For stability, since e_t is nonincreasing, $e_0 < \varepsilon$ implies $e_t < \varepsilon$ for every t. For the absence of attraction, if $\tau > 0$, then $\bar{e} > 0$, and every $e_0 \in (0, \bar{e}]$ gives $e_t \equiv e_0 \neq 0$. If $\tau = 0$, then $\bar{e} = 0$, so part (i) gives $e_t = e_0 P_t$; the argument in part (ii) yields $e_t \to 0 \iff \sum_t \lambda_t = \infty$. □

Remark (The Lyapunov Function). We have $\Delta V_t \coloneqq V_{t+1} - V_t = \frac{1}{2}(e_{t+1} + e_t)(e_{t+1} - e_t) \leq 0$. Equality holds when $e_t \leq \bar{e}$, and, if $\lambda_t > 0$, only then. When $\tau > 0$, in periods with $\lambda_t > 0$, the set on which $\Delta V = 0$ is $[0, \bar{e}]$ rather than $\{0\}$. A path starting from $e_0 > \bar{e}$ never enters $[0, \bar{e}]$, so $\Delta V_t < 0$ in periods with $\lambda_t > 0$. Nevertheless, V_t converges to

$\frac{1}{2}\bar{e}^2 > 0$, rather than zero, when $\sum_t \lambda_t = \infty$. Thus, in the argument underlying the naive view—that V decreases strictly in each period and converges to zero—it is convergence to zero that fails when the learning signal is tied to correction. In the autonomous system with $\lambda_t \equiv \lambda > 0$, the limiting set identified by LaSalle's invariance principle (LaSalle, 1976) lies within $[0, \bar{e}]$, consistently with part (ii). If $\tau = 0$ and $\sum_t \lambda_t = \infty$, however, convergence to zero does hold.

5.A.3 Proofs of Proposition 5.2 and Corollary 5.1

Proof of Proposition 5.2. Since $\Lambda_t \geq 0$, e_t is nonincreasing. By (G2), $e_{t+1} \geq \gamma e_t \geq 0$, so $e_t \in [0, e_0]$ throughout (thus $\Lambda_t(e_t)$ is defined, and the limit exists). If $e_0 \leq \bar{e}$, then (G1) gives $e_t \equiv e_0$. If $e_0 > \bar{e}$, define $T \coloneqq \inf\{t : e_t \leq \bar{e}\}$. If $T = \infty$, then $e_t > \bar{e}$ for every t, so $e_\infty \geq \bar{e} \geq \gamma\bar{e}$. If $T < \infty$, then $T \geq 1$, and (G2) implies $e_T \geq \gamma e_{T-1} > \gamma\bar{e}$. Since $e_T \leq \bar{e}$, (G1) gives $e_t = e_T$ for $t \geq T$, hence $e_\infty = e_T > \gamma\bar{e}$. □

Proof of Corollary 5.1. We have $\Lambda_t(e) = e\,(1 - \exp(-k_t\varphi_t(e))) \in [0, (1 - e^{-\bar{k}})e]$, and $e \leq \bar{e}$ implies $\varphi_t(e) = 0$, hence $\Lambda_t(e) = 0$. □

5.A.4 Proof of Proposition 5.3

(i) We have $e_{t+1} \leq (1 - \mu_t)e_t \leq e_t$. For positivity, if $e_t \leq \bar{e}$, then $e_{t+1} = (1 - \mu_t)e_t > 0$. If $e_t > \bar{e}$, then $e_{t+1} = (1 - \mu_t - \lambda_t\omega)e_t + \lambda_t\omega\bar{e}$, with both terms nonnegative. If $\lambda_t > 0$, the second term is positive ($\bar{e} > 0$); if $\lambda_t = 0$, the first term is $(1 - \mu_t)e_t > 0$.

(ii) ($\Leftarrow$) We have $e_t \leq e_0 \prod_{s<t}(1 - \mu_s) \to 0$ (Lemma 5.A.1(a)). ($\Rightarrow$) We prove the contrapositive. Suppose $\sum_t \mu_t < \infty$. If $e_t > \bar{e}$ for every t, then $\lim e_t \geq \bar{e} > 0$. Otherwise, there exists $T \coloneqq \min\{t : e_t \leq \bar{e}\}$. For $t \geq T$, we retain $e_t \leq \bar{e}$, and $e_{t+1} = (1 - \mu_t)e_t$, so $e_t = e_T \prod_{s=T}^{t-1}(1 - \mu_s) \to e_T \prod_{s \geq T}(1 - \mu_s)$. Part (i) gives $e_T > 0$, and Lemma 5.A.1(c) implies that the infinite product is positive. The limit is therefore positive. □

5.A.5 Proof of Proposition 5.4

(i) Proceed by induction on $(e_t, q_t) \in (0, e_0] \times (0, 1]$. Suppose (e_t, q_t) belongs to this set. Applying Lemma 5.1 with the substitution $w \to wq_t$ yields $0 \leq a_t^* \leq \omega(q_t)e_t < e_t$, so $e_{t+1} = e_t - \lambda_t a_t^* \in (0, e_t] \subset (0, e_0]$. Since $\omega(\cdot)$ is increasing, $a_t^* \leq \omega(1)e_0$, hence $\nu a_t^* \leq 1$. Thus $q_{t+1} \leq (1 - \delta)q_t + (1 - q_t) = 1 - \delta q_t < 1$, while $q_{t+1} \geq (1 - \delta)q_t > 0$.

(ii) Suppose, toward a contradiction, that $a_t^* > 0$ for every t. Part (i) gives $e_t > 0$, so $\sum_t \lambda_t a_t^* = e_0 - \lim_t e_t \leq e_0$. Consequently, $\sum_t a_t^* \leq e_0/\underline{\lambda} < \infty$, and $a_t^* \to 0$. Equation (5.6) gives $q_{t+1} \leq (1 - \delta)q_t + \nu a_t^*$, so $L \coloneqq \limsup_t q_t$ satisfies $L \leq (1 - \delta)L$, implying $L = 0$. Hence $\bar{e}(q_t) = \tau/(wq_t) \to \infty$. Since $e_t \leq e_0$, at some t we must have $e_t \leq \bar{e}(q_t)$, or $a_t^* = 0$, a contradiction. Therefore $T \coloneqq \min\{t : a_t^* = 0\}$ is finite. Next, we show that $a_t^* = 0$ implies $a_{t+1}^* = 0$. The equality $a_t^* = 0$ means $e_t \leq \bar{e}(q_t)$. In this case, $e_{t+1} = e_t$ and $q_{t+1} = (1 - \delta)q_t < q_t$, so $\bar{e}(q_{t+1}) > \bar{e}(q_t) \geq e_t = e_{t+1}$, implying $a_{t+1}^* = 0$.

(iii) Part (ii) gives $e_t = e_T$ for $t \geq T$. If $T \geq 1$, then $a_{T-1}^* > 0$, or $e_{T-1} > \bar{e}(q_{T-1})$. Repeating the calculation in Proposition 5.1(i) with the substitution $w \to wq_{T-1}$ yields $e_T - \bar{e}(q_{T-1}) = (1 - \lambda_{T-1}\omega(q_{T-1}))(e_{T-1} - \bar{e}(q_{T-1})) > 0$. Also, $q_{T-1} \leq 1$ implies $\bar{e}(q_{T-1}) \geq \tau/w$.

(iv) For $t \geq T$, $a_t^* = 0$, so $q_{t+1} = (1 - \delta)q_t$. □

5.A.6 Proof of Proposition 5.5

(i) Since $\lambda, \beta > 0$, we have $\Gamma_n > 1$. Monotonicity follows directly from the definition.

(ii) The two plans differ as follows. At $t = 0$, the editor's term changes from $-\frac{w}{2}e^2$ to $-\frac{w}{2}(e-\varepsilon)^2 - \tau\varepsilon - \frac{c_0}{2}\varepsilon^2$. The terms for the other $n-1$ individuals do not change. The difference is $(we-\tau)\varepsilon - \frac{w+c_0}{2}\varepsilon^2$. For $t \geq 1$, the state under plan E is $e - \lambda\varepsilon < \bar{e}$, and the private editor does not correct, so neither plan involves further correction. The loss for each of the n individuals changes from $\frac{w}{2}e^2$ to $\frac{w}{2}(e-\lambda\varepsilon)^2$, yielding a difference of $we\lambda\varepsilon - \frac{w}{2}\lambda^2\varepsilon^2$ per individual. Multiplying by $\sum_{t\geq 1}\beta^t = \beta/(1-\beta)$ and summing gives (5.9). The coefficient of the linear term is $we(1 + n\lambda\beta/(1-\beta)) - \tau = w\Gamma_n e - \tau$, which is positive when $e > e^{\text{soc}}$. The coefficient of the quadratic term is negative, so the claim follows. Since $\varepsilon < e$, we have $a_0 < e$ and $e_1 = e - \lambda\varepsilon > 0$.

(iii) For any bounded nonnegative sequence $a = (a_t)$, $e_t = e - \lambda\sum_{s<t} a_s$ is affine in a, and $|e_t|$ grows at most linearly in t, so the discounted series defining W converges absolutely. Each term is a concave function of an affine expression; hence W is concave. At $a \equiv 0$, the partial derivative with respect to a_t is the sum of the direct effect $\beta^t(we-\tau)$ and the effect through the losses of the n individuals in all periods $s > t$, namely $\sum_{s>t}\beta^s nwe\lambda = nwe\lambda\beta^{t+1}/(1-\beta)$. Thus

$$\frac{\partial W}{\partial a_t}\Big|_{a\equiv 0} = \beta^t\left(w\Gamma_n e - \tau\right)$$

The directional derivative at $a \equiv 0$ is therefore $DW(0;a) = (w\Gamma_n e - \tau)\sum_t \beta^t a_t$. If $e \leq e^{\text{soc}}$, this is nonpositive for every nonnegative a; concavity gives $W(a) \leq W(0) + DW(0;a)$, hence $W(a) \leq W(0)$. Conversely, suppose $e > e^{\text{soc}}$. The welfare difference between the plan $(\varepsilon, 0, 0, \dots)$ and $a \equiv 0$ equals the right-hand side of (5.9). (The calculation in part (ii) compares two sequences with zero correction for $t \geq 1$ and does not use $e \leq \bar{e}$.) Since the coefficient of the linear term, $w\Gamma_n e - \tau$, is positive, the difference is positive for sufficiently small $\varepsilon > 0$, so $a \equiv 0$ is not optimal. □

Appendix 5.B Details of the numerical examples and notation for Section 5

Specification The parameters in Section 5.7 are $w = 1.2$, $\tau = 0.08$, and $c_0 = 0.3$ ($\omega = 0.8$, $\bar{e} \approx 0.0667$), with $e_0 = 1$. The number of submissions, $N_t = 10 + 40/\{1 + e^{-(t-24)/8.25}\}$, illustrates an expansion in the scale of use. The learning rate $\lambda_t\omega = 1 - \exp(-\kappa\eta N_t)$ ($\kappa = 0.032$, $\eta = 0.05$) matches the strength of learning in each period to that of the path $e_t = e_0\exp(-\kappa|\mathcal{D}_t|)$, which learns independently of correction. Since $1 - \lambda_t\omega = \exp(-\kappa\eta N_t)$, P_t in (5.3) equals $\exp(-\kappa|\mathcal{D}_t|)$, yielding the expression in Section 5.7. The illustration of Proposition 5.4 uses $\delta = 0.02$, $\nu = 1$, and $q_0 = 1$ (so $\nu\omega(1)e_0 = 0.8 \leq 1$). Table 10 reports the main values, and Table 11 lists the notation.

Example for Corollary 5.1 For exponential learning with $k_t = \kappa\eta N_t \leq 0.032 \times 0.05 \times 50 = 0.08$, we have $\gamma = e^{-0.08} \approx 0.923$, giving a lower bound of at least $0.923 \times 0.0667 \approx 0.0615$. When the signal indicates whether correction occurs ($\varphi_t(e) = \mathbf{1}\{a^*(e) > 0\}$), correction stops at $t = 54$, and the error remains at approximately 0.0646. When the signal is proportional to the amount of correction ($\varphi_t(e) = a^*(e)/a^*(e_0)$), the limiting

Table 10: Main values in the numerical examples of Section 5

Quantity	Value
Lower bound $\bar{e} = \tau/w$	0.0667
Limit of V, $\frac{1}{2}\bar{e}^2$	0.00222
e_{100} on the path with learning independent of correction	0.0017
e_{100} on the path in Proposition 5.1 (correction a^*_{100})	0.0682 (0.0012)
Lower bound in Corollary 5.1, $e^{-\bar{k}}\bar{e}$ ($\bar{k} = 0.08$)	0.0615
Proposition 5.4: time T at which correction stops	72
Proposition 5.4: gap fixed at e_T	0.1057
Proposition 5.4: threshold immediately before correction stops, $\tau/(wq_{T-1})$	0.1037
Proposition 5.4: q_T, q_{100}	0.630, 0.358
Proposition 5.5: e^{soc} ($n = 1, 10, 100$; $\beta = 0.95$, $\lambda = 0.0961$)	0.0236, 0.0035, 0.00036

Table 11: Notation for Section 5

Symbol	Meaning
$e_t \geq 0$	Gap between the AI's initial output and human mathematical taste (taste-related error)
$a_t \geq 0$	Amount of human correction
$w > 0$	Weight on the loss from the gap remaining after correction
$\tau \geq 0$	Linear cost per unit of correction (effort to read and verify the passage to be corrected; marginal cost at zero correction)
$c_0 > 0$	Coefficient of the convex cost of correction
$\omega := w/(w + c_0)$	Correction response coefficient ($0 < \omega < 1$)
$\bar{e} := \tau/w$	Threshold for initiating correction (floor under correction-based learning)
$\lambda_t \in [0, 1]$	Learning rate (fraction of correction incorporated into the AI update)
Λ_t, γ, k_t, φ_t	General learning (Proposition 5.2); bound on learning per period; exponential learning (Corollary 5.1)
$\mu_t \in [0, 1)$	Rate of error reduction through learning channels other than correction (Proposition 5.3)
$q_t \in (0, 1]$, δ, ν	Acuity of human mathematical taste (sensitivity to the gap); its depreciation rate; coefficient of improvement through practice (Proposition 5.4)
n, β, Γ_n, e^{soc}	Number of users; discount factor; ratio of the private floor to the social benchmark; social benchmark (Proposition 5.5)
N_t, κ, η, $\lvert\mathcal{D}_t\rvert$	Number of submissions, learning-intensity coefficients, and cumulative data volume in the numerical example (Section 5.7)

error is exactly $\bar{e}$ (≈ 0.0667), but convergence is slow ($e_{100} \approx 0.162$). Both cases remain above this lower bound.

Example for Proposition 5.5 Using the long-run value (upper bound) of the learning rate, $\lambda \approx 0.0961$, and $\beta = 0.95$ gives $\Gamma_1 \approx 2.83$ ($e^{\text{soc}} \approx 0.0236$), $\Gamma_{10} \approx 19.3$ ($e^{\text{soc}} \approx 0.0035$), and $\Gamma_{100} \approx 184$ ($e^{\text{soc}} \approx 0.00036$). At the initial value $\lambda_0 \approx 0.0239$, we have $\Gamma_1 \approx 1.45$. In numerical calculations (a 250-period convex program), the socially optimal path starting from $\bar{e}$ approaches e^{soc} (0.0236 for $n = 1$ and 0.0035 for $n = 10$). This is a numerical observation; a full dynamic analysis with forward-looking editors and a proof of this convergence remain for future research.