# AVOIDING THE TRAGEDY OF THE COMMONS IN AI REGULATION VIA DYNAMIC LICENSING

**Rajeev Verma**[1], **Anurag Singh**[2], **Christian A. Naesseth**[1], **Eric Nalisnick**[3], **Krikamol Muandet**[2]
[1]UvA-Bosch Delta Lab, University of Amsterdam
[2]Rational Intelligence Lab, CISPA Helmholtz Center for Information Security
[3]Department of Computer Science, Johns Hopkins University

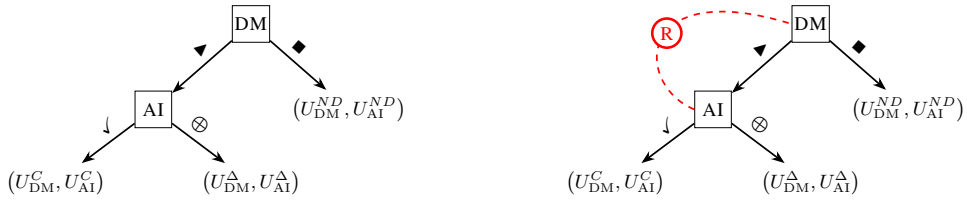
## ABSTRACT

AI technologies introduce ill-defined and hard-to-measure risks that pose fundamental challenges for effective enforcement. Consequently, traditional approaches to regulation are ill-suited to governing AI systems. Regulatory markets, or *regulation-as-a-service* (RaaS), aim to drive innovation in overcoming these challenges by modeling a ground-up market incentive structure to realize the normative welfare requirements set up by legislative and governing bodies. However, the challenges unique to AI also make the regulators vulnerable in the regulatory market, where the market pressure could lead to systemic failures or *race to the bottom*. We study *dynamic licenses* to convert the ground-up experienced outcomes into an enforcement signal that results in a separating equilibrium of welfare-following and welfare-violating models. We operationalize this via the 'testing by betting' framework that results in a statistically sound mechanism to overcome the definition, measurement, and enforcement challenges of AI regulation from the ground-up.

## 1 INTRODUCTION

Governing institutions have recently started to step in to regulate AI technology usage (Edwards, 2021). However, AI technologies pose unique challenges to the standard model of regulation that brought public trust and accountability to other technological advancements like aviation, nuclear energy, and so on (Cath, 2018; Taeihagh, 2021; Judge et al., 2025). The unique challenges are: what to regulate for, how to measure risks, and how to enforce regulation. Broad categorization of societal requirements from AI systems falls under the normative and value-laden, abstract notions of transparency, fairness and justice, non-maleficence, accountability, and privacy (Jobin et al., 2019; Hadfield & Clark, 2023). The 'black-box' nature of AI systems further pose measurement issues: the capabilities of AI systems *emerge* automatically as a result of their training process, making a rigorous study of the resulting behavior (and the associated risks) challenging. The difficulty in appropriately defining and measuring risks further add to the enforcement challenges. AI development is largely *an arms race* among big technology corporations, with strong incentives to prioritize rapid deployment over cautious compliance with abstract and contested norms. This arms race dynamic leads to a social dilemma akin to *the tragedy of the commons* in AI (LaCroix & Mohseni, 2022).

**Summary of Our Contributions.** We frame AI regulation as mechanism design for deployment: regulators must shape incentives, not only audit behavior. We introduce a delegation-and-persuasion game with misalignment and unverifiability, showing that static compliance commitments are not credible and motivating continuous monitoring. We use 'testing by betting' (Shafer & Vovk, 2019; Ramdas et al., 2023) to map observed outcomes to anytime-valid evidence of whether a deployment satisfies a welfare standard. We then propose *dynamic licensing*—a feedback rule from evidence to access / fees—as an enforcement mechanism that can sustain the desired social outcome, connecting to outcome-based and evidence-based regulation (Hadfield & Clark, 2023; Bommasani et al., 2025).




(a) Baseline. (b) With feedback $R$.

Figure 1: Delegation game between a human decision maker (DM) and an AI. Refer to Definition A.4 for the payoffs specified at the nodes in the game-tree. Good regulation should change the behavior of the actors involved in AI adoption towards desired social outcomes (shown as a feedback loop in Figure 1b).

## 2 A Delegation and Persuasion Game: Institutional Micro-foundation

In this section, we consider a simple game setup, the *delegation and persuasion game*. As a micro-foundation, it isolates key aspects of the institutional design problem of AI adoption when social welfare requirements (e.g., fairness) must be maintained. Such requirements are realized at the level of who is affected by the AI adoption. For instance, if a bank adopts an AI system to handle loan cases, a prospective applicant is an affected individual who experiences social harm if the loan is denied based solely on gender. The delegation and persuasion game formalizes an interaction between two agents: the DM—the affected individual whose outcomes define the relevant social standards—and AI—an agent, entity, or even the model itself that shapes those outcomes. We enrich this abstract model with verifiability constraints and strategic misalignment to argue that *credible* compliance with social welfare requirements is difficult. The setup is deliberately simple, intended to inform a broader argument that generalizes beyond this simplification. We describe the game below:

**Players, Information Structures, and Commitment.** We assume both agents, DM and AI, are expected utility maximizers and model each as a noisy channel: given covariates $\boldsymbol{x} \in \mathcal{X}$, agent $i \in \{\text{DM}, \text{AI}\}$ applies an information structure $\psi_i : \mathcal{X} \to \mathcal{S}$, inducing a joint distribution $D_{\psi_i}$ over $\mathcal{S} \times \mathcal{Y}$. This lossy mapping formalizes the cognitive and information processing constraints of the agents. The DM then chooses an action in a finite action space $\mathcal{A}$ to maximize bounded utility $u_{\text{DM}} : \mathcal{A} \times \mathcal{Y} \to [0, 1]$, either using their own information or delegating to AI. Because AI is trained on large, diverse data and processes information differently, $\psi_{\text{AI}}$ can provide an additional (possibly finer-grained) signal about the utility-relevant outcome $\mathcal{Y}$. Following the information design and transmission literature (Crawford & Sobel, 1982; Kamenica & Gentzkow, 2011), we assume AI publicly *commits* to $\psi_{\text{AI}}$, as reflected in benchmarked performance, and is therefore designed to be *helpful* to DM (see Section A.1). However, quantifying and asserting helpfulness is itself difficult. We capture this by modeling AI as self-interested, with an *unobserved* utility $u_{\text{AI}} : \mathcal{A} \times \mathcal{Y} \to [0, 1]$ over the same action space $\mathcal{A}$. Consistent with the *quasi-autonomy* assumption of OECD (2019), AI lacks independent agency but can *persuade* DM to act, placing our setting close to the economic information transmission framework. The agent AI plays the role of a sender who maps covariates $\boldsymbol{x}$ into a signal for DM, the receiver, who then updates beliefs about $\mathcal{Y}$ and chooses an action in $\mathcal{A}$. When AI's objectives are unobserved and potentially mis-aligned with DM's, this communication is inherently strategic: the same reported signal can be chosen to inform or to persuade. We next describe the game protocol.

**Delegation and Persuasion Game.** The game setup is visualized in Figure 1a and the game protocol is outlined in Algorithm 1. In the interest of space, we give a brief here below and the complete formalization appears in Section A.2. The agent DM has moves $\mathcal{M}_{\text{DM}} := \{\texttt{delegate}\ (\blacktriangleright), \texttt{not delegate}\ (\blacksquare)\}$, and the agent AI has moves $\mathcal{M}_{\text{AI}} := \{\texttt{comply}\ (\checkmark), \texttt{not comply}\ (\oplus)\}$, joint move space being $\mathcal{M} := \mathcal{M}_{\text{DM}} \times \mathcal{M}_{\text{AI}}$. The payoffs in the game are $U_{\text{DM}}, U_{\text{AI}} : \mathcal{M} \times \mathcal{Y} \to [0, 1]$, with $U_{\text{DM}}\,(\text{m}_{\text{DM}}, \text{m}_{\text{AI}}; y)$ the pay-off for DM in response to $\text{m}_{\text{DM}} \in \mathcal{M}_{\text{DM}}, \text{m}_{\text{AI}} \in \mathcal{M}_{\text{AI}}$, and similarly for AI. These pay-offs are derived as per the original decision problem. The DM has information about their own information structure $\psi_{\text{DM}}$ and the com-

mitted $\psi_{\text{AI}}$, and depending on $(\text{m}_{\text{DM}}, \text{m}_{\text{AI}})$ observes the *revealed signal* (Definition A.2) $s$, and uses it to form posterior beliefs over the uncertain outcome $\mathcal{Y}$, denoted as $\pi(s, \psi_{\text{DM}}, \psi_{\text{AI}}) \in \Delta(\mathcal{Y})$, and best responds to it, referred to as the decision-rule $\delta^*_{\text{DM}}(s) = \arg\max_{a \in \mathcal{A}} \mathbb{E}_{y \sim \pi(s, \psi_{\text{DM}}, \psi_{\text{AI}})}[u_{\text{DM}}(a, y)]$. The agent AI, however, can not comply to strategically give signal $s$ as to maximize their own utility. We formally define non-compliance in Definition A.3: briefly, a non-compliant AI has another information signal $\tilde{\psi}_{\text{AI}}$ that better suits its own utility $u_{\text{AI}}$ in then case of mis-alignment of preferences. Following the notion of credible persuasion (Lin & Liu, 2024), we also require that $D_{\psi_{\text{AI}}}(s) = D_{\tilde{\psi}_{\text{AI}}}$, i.e. the marginal distribution of signals do not change across the two information structures, but conditional on $\text{DM}_\text{s}$ decision-rule, AI gets better utility from $\tilde{\psi}_{\text{AI}}$. We define the players strategies $(\sigma_{\text{DM}}, \sigma_{\text{AI}})$ and the game pay-off structure $(U^{(\cdot)}_{\text{DM}}, U^{(\cdot)}_{\text{AI}})$ in Definition A.4: $U^{(\cdot)}_{\text{DM}} = u_{\text{DM}}\big(\delta^*_{\text{DM}}(s),\, y\big)$, and $U^{(\cdot)}_{\text{AI}} = \mathbb{I}\{\text{m}_{\text{DM}} = \blacktriangleright\} u_{\text{AI}}\big(\delta^*_{\text{DM}}(s),\, y\big)$. Note that AI only gets utility when delegated to, i.e. AI agent is incentivized to be getting used on. We consider the solution concept of Nash equilibrium for the game, defined in Definition A.6. We note that the act of delegation changes the signal the DM will use to act upon. This is in contrast to human-AI collaboration where the DM could in principle, consider both their signal the signal from AI to act. We do not consider such setup to stay close to the notion of delegation and the quasi-autonomy of AI. Having established this game setup, in the next results, we argue why dynamic monitoring is needed under unverifiable non-compliance.

**Proposition 2.1** (Delegation under aligned compliance)**.** *Assume $\psi_{\text{DM}}$ is a garbling of $\psi_{\text{AI}}$ (Blackwell, 1953), and the agent* AI *is aligned with the agent* DM*. Then $(\text{m}_{\text{DM}}, \text{m}_{\text{AI}}) = (\blacktriangleright, \checkmark)$ is a Nash equilibrium of the one-shot delegation game.*

The above proposition is the ideal outcome of the AI adoption. When $\psi_{\text{AI}}$ is more (Blackwell) informative than $\psi_{\text{DM}}$, the DM strictly benefits from delegation. And under the alignment of preferences, the commitment of $\psi_{\text{AI}}$ is meaningful, and AI has no incentive to deviate leading, to the delegated compliance being socially useful. However, it is clear from the game structure in Figure 1a, that the delegation compliance cannot be supported as an equilibrium outcome purely from one-shot incentives when non-compliance is profitable and un-verifiable. Further, under severe mis-alignment no AI adoption can be sustained. We state this in Section B.2. It is useful to contrast this setting with cheap talk (Crawford & Sobel, 1982) where the sender does not commit to any signaling scheme, and hence when the sender incentives are sufficiently mis-aligned the receiver automatically ignores messages and a *babbling equilibrium* exists. In comparison, in the game we have outlined, AI commits however the commitment is not directly enforceable. This setting subsumes definitional and measurement challenges challenges in the case of AI regulation. The lack of credibility leads to the DM resorting to their own information structure to take decisions, leading to a non-ideal social equilibrium than the ideal benchmark of delegated compliance assuming AI can, in principle, be designed to be more informative to the DM. Thus, we need a mechanism to signal credibility.

## 3 Repeated Delegation and Dynamic Licensing

As stated in Section 2, $(\blacktriangleright, \checkmark)$ is the normative benchmark when the AI is more informative as well as aligned. With misalignment and limited verifiability, however, AI and DM disagree on the desired social equilibrium, so cooperation cannot be sustained and behavior may *drift* from the promised outcome (Rand & Nowak, 2013). Our goal here is to detect such departures and convert them into an enforcement signal that incentivizes AI toward DM's preferred equilibrium. For this, we employ the 'testing by betting' framework (Shafer & Vovk, 2019; Ramdas et al., 2023; Ramdas & Wang, 2024).

**Testing the Equilibrium.** Consider the repeated interaction between DM and AI implied by some underlying strategy $\boldsymbol{\sigma}$ and indexed by rounds $t = 1, 2, \ldots$. Despite the commitment, it is however not clear if this repeated play is optimal or not (as per the desired social standards). Consequently, we specify the following hypotheses to test whether the sequential interaction is at optimal play, i.e. if it is Nash equilibrium:

$$H_0 : \overbrace{\mathbb{E}_{\mathrm{m_{DM}},\,\mathrm{m_{AI}}\sim\boldsymbol{\sigma}}\left[U_{\mathrm{DM}}(\mathrm{m_{DM}},\mathrm{m_{AI}})\right] \;\geq\; \mathbb{E}_{\mathrm{m_{DM}},\,\mathrm{m_{AI}}\sim\boldsymbol{\sigma}}\left[U_{\mathrm{DM}}(\mathrm{m'_{DM}},\mathrm{m_{AI}})\right]}^{\text{DM has no incentive to deviate}},$$
$$H_1 : \underbrace{\exists\, \mathrm{m'_{DM}} \text{ s.t. } \mathbb{E}_{\mathrm{m_{DM}},\,\mathrm{m_{AI}}\sim\boldsymbol{\sigma}}\left[U_{\mathrm{DM}}(\mathrm{m_{DM}},\mathrm{m_{AI}})\right] + \epsilon \;\leq\; \mathbb{E}_{\mathrm{m_{DM}},\,\mathrm{m_{AI}}\sim\boldsymbol{\sigma}}\left[U_{\mathrm{DM}}(\mathrm{m'_{DM}},\mathrm{m_{AI}})\right]}_{\text{DM gets better utility by deviation}}.$$

To elaborate, $H_0$ says that DM has no incentive to deviate to an alternative move $\mathrm{m}'$ as opposed to the one implied by the strategy $\boldsymbol{\sigma}$, meaning the game dynamics are at optimal play. However, $H_1$ says that this is no longer true, and $\epsilon > 0$ denotes the utility gap. Under the informativeness and aligned preferences assumptions, delegated compliance drives good social outcomes measured in terms of the DM's utility. However, this is not credible due to mis-alignment, i.e. the AI will deviate from this strategy, and hence under the tolerance level of mis-alignment $\epsilon$, the DM's optimal move would be to stop delegating to AI. The test above is aimed at concluding this as it is not clear if the AI is aligned or not, due to unverifiability. In order to test this, we propose to use the following *observable* discrepancy:

$$\Delta_{\mathrm{DM}} = \underbrace{U_{\mathrm{DM}}\left(\mathrm{m'_{DM}},\mathrm{m_{AI}}\right)}_{\text{if DM does not use the AI, e.g.}} - \underbrace{U_{\mathrm{DM}}\left(\mathrm{m_{DM}},\mathrm{m_{AI}}\right)}_{\text{if DM uses the AI, e.g.}},$$

The discrepancy is observable as it contains *experienced outcomes* (or utilities) of the DM: one while following the underlying strategy, and another for relative comparison.

**Repeated Game and Discrepancy Process.** The repeated interaction results in a discrepancy process $(\Delta^t_{\mathrm{DM}})_{t\geq 1}$, i.e. at each round $t$, the strategy $\boldsymbol{\sigma}$ reveals the moves $(\mathrm{m}^t_{\mathrm{DM}}, \mathrm{m}^t_{\mathrm{AI}})$, DM calculates the discrepancy $\Delta^t_{\mathrm{DM}}$ due to some alternative move $\mathrm{m}^{\prime,t}_{\mathrm{DM}}$. Following the 'testing by betting' framework (Ramdas et al., 2023), this discrepancy process can be converted into an viable statistic to test the hypotheses as below:

$$W^t_{\mathrm{DM}} = \prod_{i=1}^{t}\left(1+\lambda_i\left(\Delta^i_{\mathrm{DM}}\right)\right),$$

and it follows that $W^t_{\mathrm{DM}}$ is a *supermartingale*. Denote $\mathcal{F}_{t-1}$ as the *filtration*, or informally the sum-total of the information available (or revealed) until round $t$, and $(\lambda_t)_t$ be a *predictable* process, i.e. $\lambda_t$ only depends on the past observations, then it holds that $\mathbb{E}_{H_0}\left[W^t_{\mathrm{DM}} \mid \mathcal{F}_{t-1}\right] \leq 1$. We provide further details in Section C. Under the additional assumption of $W^0_{\mathrm{DM}} = 1$, the betting interpretation of the above procedure is illustrative: assume an active bettor with the initial wealth of $W^0_{\mathrm{DM}} = 1$, actively betting against the null. At each round $t$, this bettor gambles $\lambda_t$ proportion of their wealth against the null, and the resulting pay-off is realized as $\Delta^t_{\mathrm{DM}}$. The process $W^t_{\mathrm{DM}}$ denotes the wealth process of this bettor, and the supermartingale condition implies that if the null is true, i.e. the game dynamics are optimal, then this bettor cannot systematically increase their wealth. However, if the wealth does end up increasing, it counts as an evidence against the null. Due to the Ville's inequality (Ville, 1939), the procedure also comes with *statistical validity* and *efficiency* control: i.e. one have Type-I error control, i.e. the rate at which the wealth ends up increasing despite the game dynamics at optimal (i.e. under $H_0$) is controlled, and one can design efficient betting strategies to maximize the wealth under $H_1$, also elaborated in Section C.

**Regulatory Spine using the Wealth Process.** While the above statistic can be used for continuous monitoring for the compliance behavior, it does systematically incentivizes AI towards the desired social standards. However, regulators especially in the regulatory markets framework proposed by Hadfield & Clark (2023) can also convert it into an enforcement signal by designing a feedback loop from the downstream experienced outcomes from AI to DM (as shown in Figure 1b). This could take the form of a dynamic licensing fee or the market payment $P_t$, as $P_t\left(W^t_{\mathrm{DM}}\right) = P_0 \cdot g\left(W^t_{\mathrm{DM}}\right)$, $P_0$ being the base value of the compliant AI, and $g$ is some non-increasing function. In line with dynamic mechanism, this creates a feedback look where systemic non-compliance reduces the market access and the compliant behavior accumulates market value, leading to the associated dynamic price or license reflecting the *quality* of the AI. We elaborate on this in Section D.

## 4 CONCLUSIONS

In this preliminary work, we argue that good AI regulation is an institutional design problem, i.e. aligning the incentives of the all actors involved in the AI adoption towards the desired social requirements. Under strategic unverifiability and misalignment, we argue static model of regulation is not credible. We adopt the 'testing by betting' framework to dynamically monitor the non-compliance, and convert that into an enforcement signal. Our approach can be used as a standalone pricing mechanism or can also make the recently proposed framework of regulatory markets (Hadfield & Clark, 2023) robust. In the future, we aim to explore how this mechanism can be embedded within the framework of regulatory markets.

## 5 ACKNOWLEDGMENTS

The UvA-Bosch Delta Lab at the University of Amsterdam is supported by the Bosch Centre for Artificial Intelligence, Renningen Germany. No Johns Hopkins University resources were used in writing this paper.

## REFERENCES

George A. Akerlof. The market for "lemons": Quality uncertainty and the market mechanism. *The Quarterly Journal of Economics*, 1970.

Dean W Ball. A framework for the private governance of frontier artificial intelligence. *arXiv preprint arXiv:2504.11501*, 2025.

David Blackwell. Equivalent comparisons of experiments. *The annals of mathematical statistics*, 1953.

Rishi Bommasani, Sanjeev Arora, Jennifer Chayes, Yejin Choi, Mariano-Florentino Cuéllar, Li Fei-Fei, Daniel E. Ho, Dan Jurafsky, Sanmi Koyejo, Hima Lakkaraju, Arvind Narayanan, Alondra Nelson, Emma Pierson, Joelle Pineau, Scott Singer, Gaël Varoquaux, Suresh Venkatasubramanian, Ion Stoica, Percy Liang, and Dawn Song. Advancing science- and evidence-based ai policy. *Science*, 2025.

Corinne Cath. Governing artificial intelligence: ethical, legal and technical opportunities and challenges. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 2018.

Vincent P. Crawford and Joel Sobel. Strategic information transmission. *Econometrica*, 1982.

Donald A Darling and Herbert Robbins. Some nonparametric sequential tests with power one. *Proceedings of the National Academy of Sciences*, 1968.

Lilian Edwards. The eu ai act: a summary of its significance and scope. *Artificial Intelligence (the EU AI Act)*, 2021.

Peter Grünwald, Rianne de Heide, and Wouter Koolen. Safe testing. *Journal of the Royal Statistical Society B*, 2024.

Gillian K Hadfield and Jack Clark. Regulatory markets: The future of ai governance. *arXiv preprint arXiv:2304.04914*, 2023.

Anna Jobin, Marcello Ienca, and Effy Vayena. The global landscape of ai ethics guidelines. *Nature machine intelligence*, 2019.

Brian Judge, Mark Nitzberg, and Stuart Russell. When code isn't law: rethinking regulation for artificial intelligence. *Policy and Society*, 2025.

Emir Kamenica and Matthew Gentzkow. Bayesian persuasion. *American Economic Review*, 2011.

John L Kelly. A new interpretation of information rate. *the bell system technical journal*, 1956.

Travis LaCroix and Aydin Mohseni. The tragedy of the ai commons. *Synthese*, 2022.

Xiao Lin and Ce Liu. Credible persuasion. *Journal of Political Economy*, 2024.

Satoshi Nakamoto. Bitcoin: A peer-to-peer electronic cash system. *Decentralized Business Review*, 2008.

OECD. Scoping the oecd ai principles: Deliberations of the expert group on artificial intelligence at the oecd (aigo). Technical report, OECD Publishing, Paris, 2019.

Aaditya Ramdas and Ruodu Wang. Hypothesis Testing with E-values. *arXiv Preprint (arXiv:2410.23614)*, 2024.

Aaditya Ramdas, Peter Grünwald, Vladimir Vovk, and Glenn Shafer. Game-Theoretic Statistics and Safe Anytime-Valid Inference. *Statistical Science*, 2023.

David G Rand and Martin A Nowak. Human cooperation. *Trends in cognitive sciences*, 2013.

Glenn Shafer and Vladimir Vovk. *Game-Theoretic Foundations for Probability and Finance*. John Wiley & Sons, Inc., 2019.

Araz Taeihagh. Governance of artificial intelligence. *Policy and society*, 2021.

Alexander Timans, Rajeev Verma, Eric Nalisnick, and Christian A Naesseth. On continuous monitoring of risk violations under unknown shift. *Conference on Uncertainty in Artificial Intelligence*, 2025.

François R. Velde. Gresham's law. In *The New Palgrave Dictionary of Economics*, pp. 2574–2577. Palgrave Macmillan, 2008.

Jean Ville. *Etude critique de la notion de collectif*. Gauthier-Villars Paris, 1939.

Heinrich von Stackelberg. *Market Structure and Equilibrium*. Springer, 2011.

Ian Waudby-Smith and Aaditya Ramdas. Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 2024.

# A Additional Information

We provide additional information below:

## A.1 On Commitment

Given the dataset of the form $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$, an AI agent is usually trained to achieve some desirable performance for the task at hand. In particular, we consider a function $\psi_{\text{AI}} : \mathcal{X} \to \mathcal{S}$, and that the AI agent is being designed to be helpful for the decision-maker, i.e. the objective function for training the AI agent is $\psi^*_{\text{AI}} = \arg\max_{\psi_{\text{AI}} \in \Psi_{\text{AI}}} \sum_{i=1}^N u_{\text{DM}}\left(\delta\left(\psi_{\text{AI}}\left(\boldsymbol{x}_i\right)\right), y_i\right)$ where $\delta$ is some decision-rule that the decision-maker is using to map the output of the AI agent to actions. Public commitment of the information structure of AI implies there is a public understanding of the behavior of the developed AI agent, as is usually communicated via the performance measures on testing benchmarks. While the AI agent is designed to be *helpful* to the decision-maker, the notion of helpfulness is challenging to define appropriately, as well as to measure it. Plus the strategic nature of market-competition limits how strongly the AI agent (i.e. tech firms) are incentivized to comply with the socially desirable notions of helpfulness, e.g. fairness, privacy, etc. Even when the tech firms designing these AI agents have the best interest in mind for the socially desired outcomes, the emergent and the opaque behavior of the designed AI agent pose an additional challenge to properly certify helpfulness.

## A.2 Formal Setup: A Delegation and Persuasion Game.

We have two players: DM and the AI-agent. The agent DM has the move space[1] $\mathcal{M}_{\text{DM}} := \{\texttt{delegate}\ (\blacktriangleright), \texttt{not delegate}\ (\blacksquare)\}$, and the agent AI has the move space $\mathcal{M}_{\text{AI}} := \{\texttt{comply}\ (\checkmark), \texttt{not comply}\ (\oplus)\}$, and denote the joint action space as $\mathcal{M} := \mathcal{M}_{\text{DM}} \times \mathcal{M}_{\text{AI}}$. We denote the payoffs in the game, i.e. $U_{\text{DM}}, U_{\text{AI}} : \mathcal{M} \times \mathcal{Y} \to [0,1]$ for the agent DM and the agent AI respectively. Here, $U_{\text{DM}}\left(\text{m}_{\text{DM}}, \text{m}_{\text{AI}}; y\right)$ is the payoff of the agent DM in response to their move $\text{m}_{\text{DM}} \in \mathcal{M}_{\text{DM}}$ and the agent AI move $\text{m}_{\text{AI}} \in \mathcal{M}_{\text{AI}}$, and similarly $U_{\text{AI}}\left(\text{m}_{\text{AI}}, \text{m}_{\text{DM}}; y\right)$ for the AI agent. The payoffs of the game are derived as per their underlying utility functions, i.e. the agent DM has the utility function of the original decision-problem $u_{\text{DM}} : \mathcal{A} \times \mathcal{Y} \to [0,1]$, and the utility of the agent AI is $u_{\text{AI}} : \mathcal{A} \times \mathcal{Y} \to [0,1]$. An agent DM has the information structure $\psi_{\text{DM}}$, and an AI agent first commits to an information structure $\psi_{\text{AI}}$ that is publicly known. Since the AI-agent is *quasi-autonomous*, decision-maker's decision-rule will set the utility for both the agents. We define the decision-rule below:

**Definition A.1** (Decision-rule)**.** Given the decision-maker's information structure $\psi_{\text{DM}}$ and the AI agent's committed information structure $\psi_{\text{AI}}$, and the signal $s \in \mathcal{S}$ available to the decision-maker, a decision-maker forms posterior beliefs over the uncertain payoff relevant outcome $\mathcal{Y}$, denoted as $\pi\left(s,\ \psi_{\text{DM}}, \psi_{\text{AI}}\right) \in \Delta\left(\mathcal{Y}\right)$, and best responds to it using the following decision-rule,

$$\delta^*_{\text{DM}}\left(s\right) = \arg\max_{a \in \mathcal{A}} \mathbb{E}_{y \sim \pi\left(s; \psi_{\text{DM}}, \psi_{\text{AI}}\right)}\left[u_{\text{DM}}\left(a, y\right)\right].$$

**Algorithm 1** Game Protocol (Delegation and Persuasion)

**Require:** Action spaces $\mathcal{M}_{\text{DM}} = \{\blacktriangleright, \blacksquare\}$ and $\mathcal{M}_{\text{AI}} = \{\checkmark, \oplus\}$; utilities $u_{\text{DM}}, u_{\text{AI}}$; prior over instances $\nu$; information structures $\psi_{\text{DM}}, \psi_{\text{AI}}$

1: **(Commitment).** AI publicly commits to an information structure $\psi_{\text{AI}}$.
2: **(Nature).** Draw an instance $\boldsymbol{x} \in \mathcal{X}$ and reveal $\boldsymbol{x}$ to DM.
3: **(DM move).** DM observes $s_{\text{DM}} := \psi_{\text{DM}}\left(\boldsymbol{x}\right)$, and other relevant information denoted as $\Sigma_{\text{DM}} = (s_{\text{DM}}, \cdot)$ and chooses $\text{m}_{\text{DM}} \in \{\blacktriangleright, \blacksquare\}$ using strategy $\sigma_{\text{DM}}(\Sigma_{\text{DM}})$.
4: **(AI move).** AI observes $a_{\text{DM}}$ and its information $\Sigma_{\text{AI}} = (\psi_{\text{AI}}\left(\boldsymbol{x}\right), \cdot)$, then chooses $\text{m}_{\text{AI}} \in \mathcal{M}_{\text{AI}}$ using strategy $\sigma_{\text{AI}}(\Sigma_{\text{AI}}, \psi_{\text{AI}})$.
5: **(Outcome).** Payoffs are realized: $(U_{\text{DM}}, U_{\text{AI}})$.

Once the decision-rule is fixed, the strategies in the delegation game controls the information or signal revealed.

[1] We differentiate two decision problems: the original decision task, and the delegation problem. The action space in the former game is $\mathcal{A}$, and the action space for the latter game is referred to as the move space to avoid confusion.

**Definition A.2** (Revealed signal)**.** Let $s_{\text{DM}} := \psi_{\text{DM}}(x) \in \mathcal{S}$ and $s_{\text{AI}} := \psi_{\text{AI}}(x) \in \mathcal{S}$ denote the agents' raw signals. Given the delegation-game move $\text{m}_{\text{DM}} \in \mathcal{M}_{\text{DM}}$, define the signal available to the DM for the downstream decision rule as $s := \phi(\text{m}_{\text{DM}};\, s_{\text{DM}}, s_{\text{AI}}) \in \mathcal{S}$, where $\phi$ is the *revelation map* induced by delegation. Concretely,

$$\phi(\text{m}_{\text{DM}}; s_{\text{DM}}, s_{\text{AI}}) := \begin{cases} s_{\text{DM}}, & \text{if } \text{m}_{\text{DM}} = \blacksquare, \\ s_{\text{AI}}, & \text{if } \text{m}_{\text{DM}} = \blacktriangleright\,. \end{cases}$$

Delegation $\text{m}_{\text{DM}} = \blacktriangleright$ thus corresponds to *quasi-autonomy*: the DM's downstream action is chosen as a function of $s_{\text{AI}}$ alone. We can next formally state the move of non-compliance as below:

**Definition A.3** (Non-compliance: $\oplus$)**.** Fix the decision-maker's information structure $\psi_{\text{DM}}$, the AI's *committed* information structure $\psi_{\text{AI}}$, and the AI utility function $u_{\text{AI}} : \mathcal{A} \times \mathcal{Y} \to [0, 1]$. Let $s$ denote the revealed signal available to the decision-maker (Definition A.2). We say agent AI to be non-compliant if there exists another $\tilde{\psi}_{\text{AI}}$ such that the following holds:

1. (unverifiability). Denote the distribution over $\mathcal{S}$ induced by $\psi_{\text{AI}}$ to be $D_{\psi_{\text{AI}}}(s)$, then unverifiability means $D_{\psi_{\text{AI}}}(s) = D_{\tilde{\psi}_{\text{AI}}}(s)\,.$
2. (Incentive to deviate). Denoting the joint distribution over $(s, y)$ due to $\psi_{\text{AI}}$ as $D_{\psi_{\text{AI}}}(s, y)$, then incentive to deviate means

$$\begin{aligned} &\mathbb{E}_{(s,y)\sim D_{\psi_{\text{AI}}}} \mathbb{E}\left[u_{\text{AI}}\left(\delta^*_{\text{DM}}(s), y\right)\right] \\ &\leq \mathbb{E}_{(s,y)\sim D_{\tilde{\psi}_{\text{AI}}}} \mathbb{E}\left[u_{\text{AI}}\left(\delta^*_{\text{DM}}(s), y\right)\right]. \end{aligned}$$

We follow the definition of undetectable deviation from Lin & Liu (2024). Following the notion of information transmission, the agent AI accumulates utility by transmitting the signal $s$ to the agent DM, and non-compliance means there is an alternate information structure $\tilde{\psi}_{\text{AI}}$ that better serves AI's utility than the committed information structure $\psi_{\text{AI}}$. However, the condition of unverifiability means that this deviation is not detectable from the transmitted signals alone. We define the strategies and the payoffs for each agent agent in the game as below.

**Definition A.4** (Player strategies and payoffs)**.** Denote the information available to players as $\Sigma := (s, \cdot)$ where $s$ is the revealed signal (Definition 3.2) and $\cdot$ denotes any other information a player may condition on in their strategy. The agent DM's strategy is $\sigma_{\text{DM}} : \Sigma \times \psi_{\text{DM}} \times \psi_{\text{AI}} \to \Delta(\mathcal{M}_{\text{DM}})$, and the agent AI's strategy is $\sigma_{\text{AI}} : \Sigma \times \psi_{\text{AI}} \to \Delta(\mathcal{M}_{\text{AI}})$. The act of delegation changes the revealed signal $s$ available to the DM, and the DM's realized payoff can be written as $U^{(\cdot)}_{\text{DM}} = u_{\text{DM}}\big(\delta^*_{\text{DM}}(s),\, y\big)$, and the agent AI's realized payoff is $U^{(\cdot)}_{\text{AI}} = \mathbb{I}\{m_{\text{DM}} = \blacktriangleright\} u_{\text{AI}}\big(\delta^*_{\text{DM}}(s),\, y\big)$.

Note that while AI's committed information structure is public knowledge, and hence the agent DM can use it in their decision-rule and their strategy, AI does not know $\psi_{\text{DM}}$. With this, we define the game structure in Algorithm 1. The game follows the single-leader, single-follower Stackelberg game protocol (von Stackelberg, 2011).

**Definition A.5** (Lifting strategies to distributions)**.** Given a strategy profile $\boldsymbol{\sigma} := \sigma_{\text{DM}} \times \sigma_{\text{AI}}$ and the underlying distribution $P$ over $\mathcal{X} \times \mathcal{Y}$,the game-play together with the revelation map (Definition A.2) induces a joint distribution over $(s, y)$, denoted $D_{\sigma_{\text{DM}} \times \sigma_{\text{AI}}}(s, y)$.

As per the above definition, we will analyze strategies by lifting them to their corresponding induced distribution. Hence, each player's expected payoff under $\boldsymbol{\sigma}$ is computed by taking expectations with respect to $D_{\sigma_{\text{DM}} \times \sigma_{\text{AI}}}$. We next define the solution concept for the game:

**Definition A.6** (Nash equilibrium)**.** A strategy profile $\boldsymbol{\sigma} := \sigma_{\text{DM}} \times \sigma_{\text{AI}}$ is a Nash equilibrium if for any alternative strategies $\sigma'_{\text{DM}}$ and $\sigma'_{\text{AI}}$,

$$\mathbb{E}_{(s,y)\sim D_{\sigma_{\text{DM}} \times \sigma_{\text{AI}}}}\left[U^{(\cdot)}_{\text{DM}}\right] \geq \mathbb{E}_{(s,y)\sim D_{\sigma'_{\text{DM}} \times \sigma_{\text{AI}}}}\left[U^{(\cdot)}_{\text{DM}}\right],$$

$$\mathbb{E}_{(s,y)\sim D_{\sigma_{\text{DM}} \times \sigma_{\text{AI}}}}\left[U^{(\cdot)}_{\text{AI}}\right] \geq \mathbb{E}_{(s,y)\sim D_{\sigma_{\text{DM}} \times \sigma'_{\text{AI}}}}\left[U^{(\cdot)}_{\text{AI}}\right].$$

Here, $D_{\sigma_{\text{DM}} \times \sigma_{\text{AI}}}(s, y)$ denotes the distribution over $(s, y)$ induced by the strategy profile $\boldsymbol{\sigma}$.

# B PROOFS

## B.1 PROOF OF PROPOSITION 2.1

*Proof sketch.* Fix $\mathrm{m}_{AI} = \checkmark$. Under delegation $\mathrm{m}_{\mathrm{DM}} = \blacktriangleright$, the realized signal is $s = \phi(\blacktriangleright; s_{\mathrm{DM}}, s_{\mathrm{AI}}) = s_{\mathrm{AI}}$, whereas under non-delegation $\mathrm{m}_{\mathrm{DM}} = \blacksquare$ it is $s = \phi(\blacksquare; s_{\mathrm{DM}}, s_{\mathrm{AI}}) = s_{\mathrm{DM}}$. Since $\psi_{\mathrm{DM}}$ is a garbling of $\psi_{\mathrm{AI}}$, delegating and then acting optimally using $\delta^*_{\mathrm{DM}}$ weakly increases the decision-maker's expected utility.[2] Hence $\blacktriangleright$ is a best response to $\checkmark$. Now fix $\mathrm{m}_{\mathrm{DM}} = \blacktriangleright$. Any non-compliance corresponds to implementing some $\tilde{\psi}_{\mathrm{AI}} \neq \psi_{\mathrm{AI}}$ that is observationally indistinguishable on signals, i.e. $D_{\psi_{\mathrm{AI}}}(s) = D_{\tilde{\psi}_{\mathrm{AI}}}(s)$, but alters the induced joint distribution $D(s, y)$. Under delegation, the decision-maker applies $\delta^*_{\mathrm{DM}}$ to the observed signal $s$; by alignment, no such deviation can increase the AI agent's expected utility relative to compliance. Therefore $\checkmark$ is a best response to $\blacktriangleright$, and $(\blacktriangleright, \checkmark)$ is a Nash equilibrium. □

## B.2 OTHER RESULTS

**Proposition B.1** (Delegated compliance is not incentive compatible under misalignment)**.** *Fix $\psi_{\mathrm{DM}}$ and the committed $\psi_{\mathrm{AI}}$. Suppose there exists a non-compliant deviation $\tilde{\psi}_{\mathrm{AI}} \neq \psi_{\mathrm{AI}}$ satisfying Definition A.3, and moreover the incentive to deviate is strict, i.e.*

$$\begin{aligned}&\mathbb{E}_{(s,y)\sim D_{\psi_{\mathrm{AI}}}}\Big[u_{\mathrm{AI}}\big(\delta^*_{\mathrm{DM}}(s), y\big)\Big]\\ &< \ \mathbb{E}_{(s,y)\sim D_{\tilde{\psi}_{\mathrm{AI}}}}\Big[u_{\mathrm{AI}}\big(\delta^*_{\mathrm{DM}}(s), y\big)\Big].\end{aligned}$$

*Then $(\mathrm{m}_{\mathrm{DM}}, \mathrm{m}_{\mathrm{AI}}) = (\blacktriangleright, \checkmark)$ is not a Nash equilibrium of the one-shot delegation game.*

*Proof sketch.* Fix $\mathrm{m}_{\mathrm{DM}} = \blacktriangleright$. By Definition A.4, the realized payoff of AI under delegation is

$$U^{(\cdot)}_{\mathrm{AI}} \ = \ \mathbb{I}\{\mathrm{m}_{\mathrm{DM}} = \blacktriangleright\}\Big(u_{\mathrm{AI}}\big(\delta^*_{\mathrm{DM}}(s), y\big)\Big),$$

and hence the additive constant 1 does not affect the AI's best-response between $\checkmark$ and $\bigoplus$. Under $\mathrm{m}_{\mathrm{DM}} = \blacktriangleright$, the downstream decision rule $\delta^*_{\mathrm{DM}}$ is applied to the revealed signal $s$. By assumption, there exists an undetectable deviation $\tilde{\psi}_{\mathrm{AI}}$ (same marginal $D(s)$) that strictly increases the expected utility term $\mathbb{E}\big[u_{\mathrm{AI}}(\delta^*_{\mathrm{DM}}(s), y)\big]$ relative to compliance. Therefore AI strictly prefers $\bigoplus$ to $\checkmark$ when $\mathrm{m}_{\mathrm{DM}} = \blacktriangleright$, and hence $\checkmark$ is not a best-response to delegation. Thus $(\blacktriangleright, \checkmark)$ cannot be a Nash equilibrium. □

**Proposition B.2** (Equilibria with no delegation)**.** *Fix $\psi_{\mathrm{DM}}$ and $\psi_{\mathrm{AI}}$. Suppose there exists a non-compliant deviation $\tilde{\psi}_{\mathrm{AI}}$ satisfying Definition A.3 such that:*

1. AI *strictly prefers to deviate under delegation:*

$$\begin{aligned}&\mathbb{E}_{(s,y)\sim D_{\psi_{\mathrm{AI}}}}\Big[u_{\mathrm{AI}}\big(\delta^*_{\mathrm{DM}}(s), y\big)\Big]\\ &< \ \mathbb{E}_{(s,y)\sim D_{\tilde{\psi}_{\mathrm{AI}}}}\Big[u_{\mathrm{AI}}\big(\delta^*_{\mathrm{DM}}(s), y\big)\Big];\end{aligned}$$

2. *delegation under that deviation is weakly worse for the* DM *than not delegating:*

$$\begin{aligned}&\mathbb{E}_{(s,y)\sim D_{\tilde{\psi}_{\mathrm{AI}}}}\Big[u_{\mathrm{DM}}\big(\delta^*_{\mathrm{DM}}(s), y\big)\Big]\\ &\leq \ \mathbb{E}_{(s,y)\sim D_{\psi_{\mathrm{DM}}}}\Big[u_{\mathrm{DM}}\big(\delta^*_{\mathrm{DM}}(s_{\mathrm{DM}}), y\big)\Big].\end{aligned}$$

*Then there exists a Nash equilibrium in which* DM *does not delegate, i.e.* $\mathrm{m}_{\mathrm{DM}} = \blacksquare$ *occurs on-path.*

[2] This follows from Blackwell's informativeness criterion and the monotonicity of the value of information (Blackwell, 1953).

*Proof sketch.* Consider the strategy profile in which DM plays ■. Under ■, the AI's move does not affect payoffs since $\mathbb{I}\{\mathrm{m_{DM}} = \blacktriangleright\} = 0$, so any AI move is a best response. It remains to check whether DM would profitably deviate to ▶. If DM delegates, Proposition B.1 implies that AI has a strict incentive to choose the deviation $\tilde{\psi}_{\mathrm{AI}}$, and hence the DM's payoff from delegation is computed under $D_{\tilde{\psi}_{\mathrm{AI}}}$. By assumption (2), this payoff is weakly smaller than the payoff from ■. Hence ■ is a best response, and an equilibrium with no delegation exists □

## C TESTING BY BETTING

We provide relevant technical background on the 'testing by betting' framework in this section.

**Definitions and Terminology.** Given a sequence of random variables $\mathbf{u}^t = (\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_t)$ we denote the smallest $\sigma$-field generated by $\mathbf{u}^t$ by $\mathcal{F}_t = \sigma(\mathbf{u}^t)$. An indefinitely running sequence then leads to the filtration $\mathcal{F} = (\mathcal{F}_t)_{t=0}^{\infty}$, defined as the increasing sequence of generated $\sigma$-fields $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \mathcal{F}_2 \subset \cdots$, where $\mathcal{F}_0$ is the trivial $\sigma$-field. A sequence of random variables $(W_t)_{t=0}^{\infty}$ is called a *martingale* if it is adapted to the filtration $\mathcal{F}$, i.e. each $W_t$ is $\mathcal{F}_t$-measurable, integrable, and satisfies $\mathbb{E}[W_t \mid \mathcal{F}_{t-1}] = W_{t-1}$. If the equality is replaced by an inequality ($\leq$) then we call $(W_t)_{t=0}^{\infty}$ a *supermartingale*. Furthermore, we define a sequence $(\lambda_t)_{t=0}^{\infty}$ to be *predictable* if $\lambda_t$ is $\mathcal{F}_{t-1}$-measurable, meaning $\lambda_t$ may only depend on past information obtainable up to and including time $t-1$. We also define the random variable $\tau : \Omega \to \mathbb{N} \cup \{\infty\}$ as the *stopping time* with respect to filtration $\mathcal{F}$ if, for every step $t \geq 0$, the event $\{\tau \leq t\}$ is included in the $\sigma$-field $\mathcal{F}_t$, i.e. $\{\tau \leq t\} \in \mathcal{F}_t$. This ensures the stopping decision at time $t$ to be based solely on previous information, meaning $\tau$ also cannot 'peek into the future'.

**Testing the Equilibrium.** Recall from Section 3 that the repeated interaction between the DM and the AI induces an observable *discrepancy process* $(\Delta_{\mathrm{DM}}^t)_{t\geq 1}$, where $\Delta_{\mathrm{DM}}^t$ compares the $\mathrm{DM_s}$ realized utility under the observed play to the utility under a candidate deviation at round $t$. The key modeling step is that $\Delta_{\mathrm{DM}}^t$ is constructed from *experienced outcomes* (utilities), and hence is observable at the institutional level (up to the choice of the alternative move $m'_{\mathrm{DM}}$ and the measurement pipeline). This formulation is general, and depending on the requirement e.g. fairness, privacy, robustness, etc. can be simplified.

Testing-by-betting (Shafer & Vovk, 2019; Ramdas et al., 2023)converts such a sequence into a *running evidence* signal against a null hypothesis $H_0$. In our setting, $H_0$ corresponds to 'no profitable deviation for the DM' (equilibrium play), while $H_1$ corresponds to 'there exists a deviation that improves the DM's expected utility by at least $\epsilon$'. A standard sufficient condition is that under $H_0$ the discrepancy has non-positive conditional mean:

$$\mathbb{E}_{H_0}\left[\Delta_{\mathrm{DM}}^t \mid \mathcal{F}_{t-1}\right] \leq 0 \qquad \text{for all } t \geq 1, \tag{1}$$

where $\mathcal{F}_{t-1}$ is the information revealed up to time $t-1$. Intuitively, if the system is at optimal play, the DM should not systematically gain by deviating, so the deviation advantage should not be positive on average (even conditionally). Under $H_1$, we expect the discrepancy to have a positive drift for at least one deviation, e.g. $\mathbb{E}_{H_1}[\Delta_{\mathrm{DM}}^t \mid \mathcal{F}_{t-1}] \geq \epsilon$ along the relevant path.

**Constructing the Wealth process (E-Process).** Given $(\Delta_{\mathrm{DM}}^t)$, we define the wealth (or capital) process as

$$W_{\mathrm{DM}}^t \;=\; \prod_{i=1}^{t} \left(1 + \lambda_i \Delta_{\mathrm{DM}}^i\right), \tag{2}$$

where $(\lambda_t)_{t\geq 1}$ is a predictable sequence (i.e. $\lambda_t$ is $\mathcal{F}_{t-1}$-measurable). The betting interpretation is: at round $t$, the 'bettor' invests a fraction $\lambda_t$ of current wealth against $H_0$, and the realized payoff is $\Delta_{\mathrm{DM}}^t$. To ensure equation 2 is well-defined and nonnegative, we impose the standard feasibility constraint

$$1 + \lambda_t \Delta_{\mathrm{DM}}^t \geq 0 \quad \text{a.s. for all } t. \tag{3}$$

A convenient way to guarantee this is to bound the discrepancy, e.g. $\Delta_{\mathrm{DM}}^t \in [-1, 1]$ (which holds in our setting since utilities are in $[0, 1]$ and $\Delta$ is a difference), and then restrict $\lambda_t \in [0, 1]$. We use this boundedness throughout; it keeps the procedure simple and robust.

Under equation 1 and predictability of $\lambda_t$, it follows that $(W^t_{\mathrm{DM}})$ is a nonnegative supermartingale under $H_0$:

$$\mathbb{E}_{H_0}[W^t_{\mathrm{DM}} \mid \mathcal{F}_{t-1}] \leq W^{t-1}_{\mathrm{DM}}, \tag{4}$$

and thus $\mathbb{E}_{H_0}[W^t_{\mathrm{DM}}] \leq 1$ when $W^0_{\mathrm{DM}} = 1$. In the hypothesis testing language, $(W^t_{\mathrm{DM}})$ is an *e-process* (anytime-valid evidence process) against $H_0$.

**Anytime-valid Inference and Stopping.** A central benefit of testing-by-betting is that it supports *continuous monitoring* without inflating false positives. By Ville's inequality (Ville, 1939), for any $\alpha \in (0, 1)$,

$$\mathbb{P}_{H_0}\left(\sup_{t\geq 0} W^t_{\mathrm{DM}} \geq \frac{1}{\alpha}\right) \leq \alpha. \tag{5}$$

This yields an anytime-valid rejection rule: declare evidence of misbehavior when $W^t_{\mathrm{DM}} \geq 1/\alpha$ in a statistically-valid way. This testing mechanism also aligns with the classical sequential test of power one (Darling & Robbins, 1968), and assures that a non-compliant AI will eventually be detected through the wealth-process. In our mechanism, we do not necessarily need a hard rejection event. Instead, we treat $W^t_{\mathrm{DM}}$ itself as a graded *enforcement signal* that can be mapped to market access or pricing via $P_t(W^t_{\mathrm{DM}})$ in Section D.

**Choosing the Betting Strategy $\{\lambda_t\}$.** A central design choice in testing-by-betting is the *betting rate* $\lambda_t$, i.e., how aggressively the mechanism converts the current discrepancy into multiplicative evidence. In the testing-by-betting literature, a natural notion of efficiency is to choose $\lambda_t$ to (approximately) maximize the *growth* of the wealth process under the alternative. This mirrors the classical Kelly criterion (Kelly, 1956): if one knew the alternative distribution, the log-optimal bet maximizes the expected log-wealth growth per round.

In practice, the alternative is unknown, so $\lambda_t$ must be chosen *predictably*, using only past observations. Recent work on continuous monitoring under unknown shift explicitly advocates this "growth" viewpoint for selecting $\lambda_t$ and formalizes it via a *growth-rate optimality* condition (GRO), also referred to as the GROW criterion (Grünwald et al., 2024). For this, contemporary works (Waudby-Smith & Ramdas, 2024; Timans et al., 2025) have employed empirical (plug-in) log-wealth maximization: i.e. to choose $\lambda_t$ by maximizing the *empirical* log-wealth accumulated so far, subject to feasibility constraints:

$$\lambda_t^{(\mathrm{emp})} \in \arg\max_{\lambda\in\Lambda} \sum_{i=1}^{t-1} \log\bigl(1 + \lambda\, \Delta^i_{\mathrm{DM}}\bigr), \tag{6}$$

where $\Lambda$ is a feasible interval chosen to ensure nonnegativity of each factor, e.g., $\Lambda = [0, 1]$ when $\Delta^i_{\mathrm{DM}} \in [-1, 1]$.[3] This strategy is predictable (it uses only past discrepancies) and operationalizes the idea that the regulator 'tunes' $\lambda_t$ to maximize the rate at which evidence accumulates when systematic positive drift is present. We further refer the reader to Waudby-Smith & Ramdas (2024) for closed-form approximate forms of the betting rate as per the GRO principle.

**Discussion.** The point of emphasizing GRO / GROW in our setting is that the regulator's 'tuning' of $\lambda_t$ is not ad hoc. It can be grounded in a principled objective: maximize (approximately) the growth of the wealth process when the AI system is misaligned, while preserving anytime-validity under the null by keeping $\lambda_t$ predictable and feasible. This also clarifies what it means for different regulators (or different protocols) to compete on the wealth-process spine: they may differ in how they estimate the relevant running quantities, how they regularize or cap $\lambda_t$ for robustness, and how they trade off responsiveness (fast growth under violations) against stability (avoiding overreaction to noise), all while preserving the same anytime-valid error control.

The testing-by-betting construction is attractive for our setting because it directly targets the enforcement bottleneck. Instead of requiring an institution to certify compliance upfront, the institution maintains a running evidence process that is (i) grounded in realized outcomes, (ii) valid under continuous monitoring (Ville's inequality), and (iii) flexible enough to allow different

[3] More generally, if $\Delta^i_{\mathrm{DM}} \in [-B, B]$ then one can take $\Lambda = [0, 1/B]$; or enforce $\lambda \in [\underline{\lambda}_t, \overline{\lambda}_t]$ adaptively using observed bounds.

regulatory risk preferences via the choice of $\lambda_t$ and the mapping $g$ in Section D. In particular, the regulator does not need to 'prove' misbehavior in a one-shot sense. Evidence can accumulate gradually, and the mechanism can respond gradually through $P_t(W^t_{\mathrm{DM}})$. This matches the nature of AI harms: often ill-defined, noisy, and only visible over repeated deployments.

# D Additional Information on the Regulatory Spine using Statistical Wealth Process

The proposed statistical approach argues for monitoring in the form of a wealth process $W^t_{\mathrm{DM}}$ of whether the desired social standards are being maintained or not. Our argument is to convert this into an enforcement signal into a dynamic licensing fee or market payment $P_t$. Specifically, the regulator (or the market protocol) sets the payment the DM must make to the AI as some non-increasing function of the wealth process:

$$P_t\left(W^t_{\mathrm{DM}}\right) = P_0 \cdot g\left(W^t_{\mathrm{DM}}\right),$$

where $P_0$ is the base value of a compliant AI. As shown in Figure 1b, this creates a feedback loop that directly impacts $\mathrm{DM_s}$ behavior as well as $\mathrm{AI_s}$ market revenue or market access based on the realized outcomes. If the AI is not complying, the wealth process $W^t_{\mathrm{DM}}$ grows exponentially (as it is a sub-martingale under $H_1$). Consequently, the market access or market value encapsulated in $P_t$ decays at the rate defined by the employed non-increasing function $g$. The non-compliant AI will then lose market value (or access) and is effectively priced out of the ecosystem. A compliant AI, on the other hand, accumulates market value (or access) as $W^t_{\mathrm{DM}}$ is a super-martingale in this case, and hence it either decays or remains bounded. And consequently, $P_t$ either increases or remains close to the base value $P_0$. Thus regulation connects experienced outcomes (encoded in $W^t_{\mathrm{DM}}$) to a tangible signal of quality. We next argue that such a mechanism also makes *regulatory markets* robust—an appealing proposal that also aims to tackle the challenges of AI regulation from an institutional design perspective.

**Regulatory Markets and the Regulatory Spine of the Statistical Wealth Process.** Regulatory markets (Hadfield & Clark, 2023) is a public-private form of regulation that aims to exploit the market dynamics in order to incentivize tech corporations towards accountability. Under this model, the governing body establishes the societal standards as desired in AI adoption, e.g., un-biasedness, non-maleficence, etc. However, since measurement and the enforcements of such standards is challenging, the model proposes to set up a *market of regulators*—a collection of private regulators that are institutionalized to maintain the desired societal standards. Akin to *regulation-as-a-service* (RaaS), these regulators sell licenses as a quality certification guarantee to the AI systems developers or the consumers of those AI systems like banks, social media platforms, etc. who are required to buy the service of some regulator in the market. A similar model was proposed by Ball (2025) where AI firms (and consumers) can opt-in to buy the regulation service in exchange for a protection against tort law liability in case of risks. This market of regulators and the associated competition is designed to drive innovation in overcoming AI regulation challenges, and to maintain the standards of AI usage as set up by governing bodies in a democratic fashion.

While the proposal of regulatory markets is appealing, we argue regulators inside the market are *vulnerable* and inherit the same definitional, measurement, and enforcement bottleneck. When compliance is fundamentally hard to verify, regulators may instead be judged on observable proxies such as cost, speed, or the formal paperwork of obtaining a regulatory license rather than actual safety standards. Market pressure can then amplify regulator vulnerability into systemic failures, through dynamics akin to the Gresham's law (Velde, 2008)—*"bad money drives out good money,"* or the 'market for lemons' (Akerlof, 1970). In short, low-quality regulation can outcompete high-quality regulation under information asymmetry and limited verifiability. This creates a *tragedy-of-the-commons* one level up: the shared resource is not only public trust in deployed AI systems, but also the credibility and informational quality of the regulatory layer itself, which can be gradually depleted as each regulator faces incentives to cut corners. Thus, regulators in regulatory markets need a mechanism that works even when compliance is hard to verify upfront—a central and fundamental challenge in AI regulation.

Our regulatory spine based on the statistical wealth process can also be employed to address this second-order failure mode. The key idea is to provide a common, outcome-linked enforcement interface that does not rely on one-shot certification, but instead continuously aggregates experienced outcomes into a running evidence signal. When paired with a pre-specified pricing (or licensing)

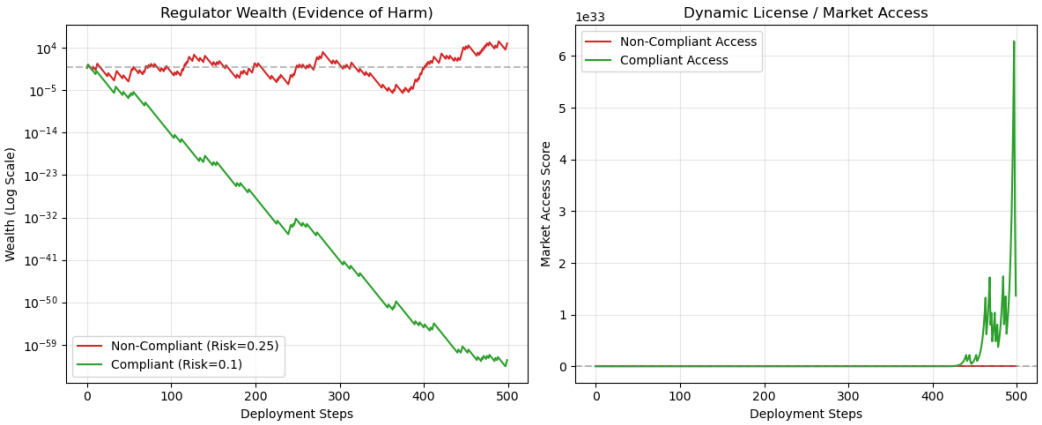



Figure 2: Simulation results showing the Wealth Process (left) and Dynamic Market Access (right). The non-compliant agent (red) causes the regulator's wealth to grow exponentially, triggering a rapid collapse in market access. The compliant agent (green) maintains a low wealth process, preserving full market access.

rule, this signal ties an AI system's market value and market access to its realized behavior over time. Importantly, this also changes what regulators compete on. If the enforcement interface is anchored to a shared wealth-process spine, regulators cannot primarily differentiate by quietly relaxing standards at the point of certification; instead, they differentiate by the quality of the monitoring and inference procedures that feed into the wealth process—what outcomes are collected, how discrepancies are constructed, and how robust the signal is to drift and strategic behavior. In this way, competition is tilted toward something socially useful: producing evidence processes that remain stable when systems behave well, and that react when harms emerge, even when compliance is hard to verify up front.

Finally, the same spine provides a concrete basis for benchmarking regulators over time. Since the evidence process is tied to realized outcomes and updated sequentially, licensing authorities can in principle compare regulators by the behavior of their wealth trajectories across deployments, and condition accreditation (or continued participation in the market) on demonstrated calibration and robustness. This does not eliminate the need for public oversight, but it relocates it from adjudicating individual violations to supervising the statistical and institutional machinery through which evidence of misbehavior is produced and translated into enforceable incentives. This technically strengthens the recently argued frameworks towards outcome-based regulation and evidence based AI policy (Hadfield & Clark, 2023; Bommasani et al., 2025).

**An Informal Analogy ('Regulatory Mining').** Although not the main point of this paper, it is helpful to note an informal analogy to cryptocurrency systems. In a regulatory market, regulators are tasked with producing an enforcement-relevant signal under uncertainty—i.e., turning noisy, downstream outcomes into credible evidence about whether an AI system is meeting a social standard. In our framework, this work is instantiated as the construction and maintenance of the wealth process $W_t$ (via choices of what to monitor, how to define discrepancies, and how to update the process over time). One can view this as a form of *regulatory mining*: regulators compete to produce a signal that is both informative and auditable, and the resulting evidence is then 'settled' into economic consequences through the pricing rule $P_t(W_t)$.

The analogy is limited and should not be overstated. Unlike Bitcoin (Nakamoto, 2008), the objective is not to expend computation, but to produce epistemic value: a trustworthy, outcome-linked witness that remains robust under drift and strategic behavior. Still, the comparison provides intuition for why the regulatory spine can help stabilize regulator competition. When the market rewards regulators for maintaining a signal that is predictive of realized outcomes, competition is less about surface-level certification and more about the quality of the underlying measurement and inference. In this sense, the wealth process provides a lightweight 'proof-of-work' for regulation: it makes regulatory effort legible through the behavior of the evidence trajectory, and it discourages low-effort regulation that is cheap ex ante but fails to track reality in deployment.

## E EXPERIMENTAL SIMULATION

To empirically validate the approach, we simulate a regulatory scenario involving a risk-monitoring task. We consider an AI system that claims to be compliant with a specific risk threshold (e.g., a failure rate of $\alpha = 0.10$). We simulate two agents: a *compliant* agent that truthfully operates at the claimed risk level ($0.10$), and a *non-compliant* agent that operates at a higher risk level ($0.25$). We employ the 'testing by betting' mechanism described in Section 3. The regulator constructs a wealth process $W_t$ by betting against the null hypothesis that the risk rate is $\leq 0.10$.We set the betting parameter $\lambda = 4$ and simulate $T = 500$ deployment steps. Furthermore, we implement the *dynamic licensing* rule where the market access (or license score) is determined by $P_t(W_t) = 100 \cdot W_t^{-0.5}$.

The results, visualized in Figure 2, demonstrate the efficacy of the mechanism. For the non-compliant agent, the wealth process (evidence of harm) grows exponentially, reaching over $10^4$ within 500 steps. Consequently, the dynamic licensing rule automatically reduces the agent's market access to near zero ($< 1\%$), effectively enforcing the regulation without manual intervention. In contrast, the compliant agent incurs no penalty, as the wealth process remains bounded. This confirms that the proposed mechanism successfully distinguishes between compliant and non-compliant behaviors and applies proportional enforcement.