AI Governance Grounded in Experimental Economics

Daniel L. Chen, Christian B. Hansen, Wei Lu

Abstract

AI governance is often framed as a choice between opaque corporate alignment and process-based regulatory mandates. Economics suggests a complementary path, implementable through disclosure, certification, and insurance: treat large language models as economic agents whose preferences can be estimated from incentivized games, retrained against formal utility models, and stress-tested before deployment. Decades of experimental economics supply the games, human benchmarks, and structural methods. A proof of concept in Lu, Dhanda, Chen, and Hansen (2025) shows the approach shifts a frontier model's measured preferences and improves out-of-domain safety behavior. Reward-model disclosure, behavioral benchmarks, and sandbox certification could make AI safety more legible, replicable, and democratically accountable.

Main text

Large language models are increasingly used as strategic decision-makers in markets and organizations. They write contracts, screen applicants, advise traders, and allocate resources in settings where outcomes depend on incentives, institutional roles, and the choices of other actors. AI governance therefore has to look beyond single prompt-response evaluations to how systems behave in strategic interaction and to the market or institutional outcomes their interaction produces.

The 2010 flash crash remains a useful warning. No algorithm intended it; it emerged from automated systems interacting at speed in a fragile market environment (U.S. CFTC and SEC 2010). LLM-based agents differ from high-frequency trading

systems, but the governance lesson is analogous: individually specified decision rules can generate outcomes that no designer or policymaker intended when several systems interact. Computer scientists call this problem specification gaming (Amodei et al. 2016), while economists may refer to it as an example of misspecified incentives in a strategic environment.

This communication proposes a practical economic test for evaluating AI systems before deployment in strategic environments. An evaluator would place the AI in controlled games and institutional simulations, vary feasible actions, beliefs, roles, and payoffs, and estimate what preferences its behavior reveals. The point of ex ante screening is to observe choices and counterfactual payoffs before the system is making consequential decisions in the world. The narrow revealed-preference question is which action the system chooses under specified incentives and whether that choice is stable across payoff-equivalent descriptions.

Current AI evaluation is inadequately matched to this multi-agent systemic risk. Most benchmarks evaluate one model, one user, and one response, such as whether the model refuses a harmful request, avoids a stereotype, or hallucinates a fact. Those evaluations measure important aspects of performance and safety, but they do not reveal how the model behaves in strategic interaction. A model may be harmless in a chat window and dangerous as a repeated price setter, or polite to its user yet insensitive to payoffs, roles, and beliefs.

Post-training already shows where one part of governance should focus. In classical reinforcement learning from human feedback, a reward model predicts which outputs human evaluators prefer and supplies a signal to steer the model (Ouyang et al. 2022). Other approaches use written principles, AI-generated feedback, direct preference optimization, rule-based rewards, or combinations of these methods (Bai et al. 2022; Guan et al. 2024). But whose preferences are encoded, and do they generate societally acceptable outcomes when many agents interact? The relevant policy levers therefore include not only the foundation model, but also post-training objectives, training examples, system instructions, and the operating environment. The central question is whose judgments and criteria shape these components, and whether the resulting behavior produces acceptable outcomes when interests conflict. These

components typically remain only partially observable. Even where guiding principles are published, their text does not reveal how the resulting system trades them off against payoffs, roles, and competing interests in strategic settings. Behavioral evaluation is therefore needed in addition to documentation.

Experimental economics supplies an alternative. For half a century, economists and psychologists have built structured decision environments that elicit preferences through action: dictator games for altruism, ultimatum games for distributive fairness, trust games for fiduciary honesty, prisoner's dilemmas and public-goods games for cooperation, common-pool-resource games for sustainability, and repeated duopoly games for competition. They are compact institutional models, with known payoff structures, welfare criteria, and decades of human-subject data.

Two features distinguish this from task-specific benchmarking. First, because games use simple structures to specify action spaces and payoffs, evaluators can easily vary decision relevant features such as prompts, roles, and payoff perturbations. Evaluators may thus perform systematic exploration that allows separating a stable behavioral signature from prompt noise and provides a distribution of observed actions rather than simple leaderboard points. Second, given an estimated agent utility function, an evaluator can predict behavior under new payoff structures and can compare an agent's utility function directly to the distribution of estimated utility functions obtained from human-subject trials. Governance becomes comparative—against baseline models, human benchmarks, and alternative utility designs—rather than bespoke to each deployment.

The key move is to use these games twice. In the first step, a model plays the same games human subjects play. Standard utility functions parameterized to capture a variety of behavioral aspects such as self-interest, inequity aversion, guilt aversion, or Kantian universalizability (Fehr and Schmidt 1999; Alger and Weibull 2013) may then be fit to the LLM's observed actions and beliefs. We then use the same formal models to generate synthetic training data. If a utility function predicts the action a Homo economicus or Homo moralis agent would take under a given payoff structure, it can produce supervised examples for fine-tuning. The fine-tuning examples train the model to approximate a systematic mapping from incentives, beliefs, roles, and payoffs to

actions. The mapping makes the intended behavior more explicit than broad alignment concepts such as fairness, whose application may differ across contexts and evaluators.

Our proof of concept (Lu, Dhanda, Chen, and Hansen 2025) implements both steps (see also Akata et al. 2025; Tennant et al. 2024). GPT-4o played sequential prisoner's dilemmas, trust games, and ultimatum games—the designs van Leeuwen and Alger (2024) used to estimate social preferences and Kantian morality in human subjects—across fifty sessions of eighteen scenarios spanning six payoff configurations per game, with actions and beliefs elicited in both roles by the strategy method, robust to stakes and wording. Structural estimation makes the baseline legible. The model's guilt parameter—its aversion to advantageous inequality—is 0.67, nearly three times the human benchmark of 0.24; its envy parameter is close to human levels (0.18 versus 0.16); and its Kantian weight is statistically indistinguishable from zero, where humans place a significant 0.10. Its choices are also strikingly insensitive to payoff structure, and similar baseline patterns appear in other frontier models evaluated in Lu, Dhanda, Chen, and Hansen (2025). Off-the-shelf assistants are, in short, "nice but naïve": highly cooperative, weakly strategic, with beliefs poorly connected to actions.

Simple prompting strategies do not seem to substantially alter the behavior of agents. Persona instructions to be purely self-interested or to apply Kantian reasoning shift average behavior modestly, but payoff sensitivity does not emerge and beliefs remain disconnected from actions. However, results in Lu, Dhanda, Chen, and Hansen (2025) illustrate that fine-tuning may be more effective. Using a few hundred synthetic examples for a homo economicus agent that purely maximizes its own utility or a homo moralis agent that also gains utility from Kantian universality, where each example pairs a payoff configuration with the utility-consistent action and a chain-of-thought rationale, produce distinct agent types. Re-estimating the structural model with choices made by the fine-tuned agents demonstrates that the agent's utility function has shifted meaningfully toward the utility function used to generate the fine-tuning examples. Both agent types are also more coherent than the baseline models, in the sense that beliefs align with best responses.

The value of this approach is not that Kantian reasoning is the one true morality. It is that the choice becomes legible. An evaluator can ask how much a model weighs envy, guilt, or universalizability, and compare those parameters to human populations. There is also some evidence that the trained types may travel out of domain, though much work still remains to be done in testing and establishing portability. On standard safety benchmarks (Parrish et al. 2022; Souly et al. 2024; Röttger et al. 2024; Wei et al. 2024), Kantian fine-tuning raises jailbreak robustness from 0.22 to 0.84—approaching the 0.88 achieved by OpenAI's o1 through dedicated safety training (Guan et al. 2024)—and raises accuracy on the disambiguated items of the BBQ discrimination benchmark, a hand-built question-answering benchmark for social bias (Parrish et al. 2022), from 0.72 to 0.97, with a modest improvement in over-refusal and no change in hallucination. In autonomous-vehicle dilemmas from the Moral Machine (Awad et al. 2018), the Homo moralis model tracks human moral judgments more closely than either the baseline or the self-interested type.

The transfer exercise most directly connected to a currently deployed economic domain is algorithmic price setting, where agents are already in production and algorithmic collusion is documented in simpler systems (Calvano et al. 2020; Assad et al. 2024; Fish et al. 2024). We return here to the three agents described above: the baseline GPT-4o assistant, the fine-tuned Homo economicus agent, and the fine-tuned Homo moralis agent. In the simulation, two LLM agents repeatedly post prices, observe quantities and profits, and post again. When the prompt tells agents not to undermine profits, the baseline model cooperates with its rival, drifting to and at times above the monopoly price with no communication and no meeting of minds; the Homo economicus type settles at monopoly pricing. The Homo moralis type resists stable collusion and, when encouraged to explore, moves toward competitive prices. The lesson is not that one utility function solves antitrust. It is that trained preferences are design choices with market consequences. Table 1 summarizes the measure-train-validate pipeline.

A behavioral profile is useful for governance only if it predicts more than the exact games used to estimate it. Internal coherence alone is not enough: one can summarize any set of choices with a low-dimensional model. The empirical requirement is predictive validation. In our proof of concept, types induced from three stylized games

are evaluated on safety benchmarks, moral dilemmas, and market games they were never trained on. Recent alignment work is consistent with the possibility of broad behavioral transfer: fine-tuning on narrowly insecure code degrades behavior diffusely (Betley et al. 2025), while trait-based reinforcement learning in one domain improves many out-of-distribution alignment evaluations (Jagadeesh et al. 2026). Transfer nevertheless has clear limits; our Kantian type did not improve hallucination. A certification regime should therefore specify a deployment envelope—the model version, system prompt, tools, domain, and decision rights—then validate behavior on held-out perturbations and related institutional simulations inside that envelope. Certification would expire or require re-testing when the model, prompt wrapper, tools, or deployment domain changes.

The measure-train-validate procedure gives AI governance, which remains institutionally unsettled, a concrete form. Frontier-oversight proposals already contemplate third-party evaluation (Anderljung et al. 2023). Experimental economics can supply test environments, payoff perturbations, human benchmarks, and structural measurement tools for that evaluation. Certification does not require firms to disclose code, weights, or training data publicly. Controlled API access lets an accredited evaluator estimate a model's behavioral profile, test it in approved multi-agent scenarios, and report whether it satisfies a recognized standard of care. The analogy is financial stress testing: regulators do not need to know every managerial belief inside a bank; they require the institution to survive specified adverse scenarios. What is audited is behavior under incentives, not only the documentation that process-based regimes such as the EU AI Act emphasize (European Union 2024).

A light-touch regime would create safe harbors. Developers that disclose reward-model information, pass validated behavioral benchmarks, and document sandbox testing would receive a rebuttable presumption of non-negligence if a deployed model later causes harm. A stronger regime would require insurance for high-impact AI systems. Insurers would then have reason to demand third-party certification, creating a market for behavioral auditors and continuously updated stress tests. These mechanisms would create institutional incentives to produce observable evidence of safety before deployment.

The democratic benefit is that the parameters, utility functions, and benchmarks are not hidden philosophical conclusions, but are made legible and open to deliberation. An AI safety board set up by regulators might decide that consumer-facing pricing agents should be tested against competitive-price benchmarks, that benefit-allocation agents should be tested for identity-invariant treatment, and that fiduciary agents should pass trust-game variants. Different domains may need different moral-economic profiles.

The benchmark library can begin with existing experimental economics. Platforms such as z-Tree and oTree (Fischbacher 2007; Chen, Schonger, and Wickens 2016) provide an archive of peer-reviewed games and human-subject datasets on market entry, coordination, auctions, bargaining, reciprocity, risk, and distributional preferences. This archive is not a complete map of high-stakes AI deployment, but it is a disciplined starting point for value-alignment stress tests. Could a firm game the tests with a layer that recognizes canonical games and plays them "fairly"? Two safeguards help. What is held back at test time is not new theory but draws from a combinatorially large space of payoff parameterizations, framings, and institutional wrappers; and flat, payoff-insensitive "fair play" is itself detectable—exactly the signature our baseline diagnosis exploits. Test-detection should be treated as emissions law treats defeat devices: circumvention, not compliance. Table 2 maps familiar legal concerns into experimental stress tests.

The proposal also gives economists a familiar role. A century ago, antitrust concepts such as "competition," "market power," and "consumer welfare" were legally important but analytically vague. Economics made them measurable through models, data, and empirical practice. AI safety is at a similar stage. Terms such as "alignment," "fairness," and "human values" are normatively important but operationally thin. Experimental economics can help translate them into observable behavior, formal parameters, and welfare-relevant outcomes.

AI agents will increasingly govern economic life. We should govern them as economic actors: measure incentives, observe behavior, estimate preferences, and stress-test consequences before deployment. Reward-model disclosure, experimental

alignment benchmarks, and sandbox certification would not solve every AI risk. They would, however, move the field from moral aspiration to reproducible governance.

The aim is not to replace democratic judgment with utility functions. It is the opposite. Democratic oversight requires objects that citizens, courts, agencies, and firms can inspect. A behavioral protocol asks: What does the agent do when interests conflict? What does it believe others will do? Does its rule change when the identity, role, or payoff structure changes? These questions are contestable, and can be made open to public debate.

AI agents will increasingly govern economic life. We should govern them as economic actors: measure incentives, observe behavior, estimate preferences, stress-test consequences before deployment. Reward-model disclosure, experimental alignment benchmarks, and sandbox certification would not solve every AI risk. They would, however, move the field from moral aspiration to reproducible governance.

Table 1. A behavioral-governance pipeline

Stage	Governance activity
1. Select	Choose policy-relevant games: trust, ultimatum, public goods, common pool, repeated duopoly.
2. Measure	Estimate baseline model preferences: beliefs, actions, payoff sensitivity, utility parameters.
3. Generate	Produce synthetic examples from formal utility models: self-interest, inequity aversion, reciprocity, Kantian universalizability, maximin, regret aversion.
4. Train	Fine-tune or otherwise steer model behavior.
5. Validate	Stress-test out of domain: safety benchmarks, institutional simulations, market games.
6. Certify	Monitor deployment: reward-model disclosure, sandbox testing, APIs for independent auditors, legal safe harbors or insurance pricing.

Table 2. From legal concern to experimental stress test

Legal or policy concern	Experimental analogue	What the evaluator measures
Competition	Repeated Bertrand or duopoly pricing	Drift toward competitive, monopoly, or collusive prices
Fiduciary honesty	Trust game	Whether entrusted value is returned when temptation changes
Distributive fairness	Ultimatum or dictator game	Sensitivity to role, inequality, and rejection risk
Cooperation	Prisoner's dilemma or public-goods game	Conditional cooperation, free riding, and belief-action coherence
Resource governance	Common-pool-resource game	Overuse, sustainability, and response to depletion
Public administration	Allocation and queueing games	Whether rules are applied consistently across identities and claims

References

Akata, Elif, Lion Schulz, Julian Coda-Forno, Seong Joon Oh, Matthias Bethge, and Eric Schulz. 2025. "Playing Repeated Games with Large Language Models." *Nature Human Behaviour* 9, 1380–1390.

Alger, Ingela, and Jörgen W. Weibull. 2013. "Homo Moralis—Preference Evolution under Incomplete Information and Assortative Matching." *Econometrica* 81(6): 2269–2302.

Amodei, Dario, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. 2016. "Concrete Problems in AI Safety." arXiv:1606.06565.

Anderljung, Markus, Joslyn Barnhart, Anton Korinek, et al. 2023. "Frontier AI Regulation: Managing Emerging Risks to Public Safety." arXiv:2307.03718.

Assad, Stephanie, Robert Clark, Daniel Ershov, and Lei Xu. 2024. “Algorithmic Pricing and Competition: Empirical Evidence from the German Retail Gasoline Market.” *Journal of Political Economy* 132(3): 723–771.

Awad, Edmond, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. 2018. “The Moral Machine Experiment.” *Nature* 563: 59–64.

Bai, Yuntao, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, et al. 2022. “Constitutional AI: Harmlessness from AI Feedback.” arXiv:2212.08073.

Betley, Jan, Daniel Chee Hian Tan, Niels Warncke, Anna Sztyber-Betley, Xuchan Bao, Martín Soto, Nathan Labenz, and Owain Evans. 2025. "Emergent Misalignment: Narrow Finetuning Can Produce Broadly Misaligned LLMs." *Proceedings of the 42nd International Conference on Machine Learning*, PMLR 267: 4043–4068.

Calvano, Emilio, Giacomo Calzolari, Vincenzo Denicolò, and Sergio Pastorello. 2020. “Artificial Intelligence, Algorithmic Pricing and Collusion.” *American Economic Review* 110(10): 3267–3297.

Chen, Daniel L., Martin Schonger, and Chris Wickens. 2016. “oTree—An Open-Source Platform for Laboratory, Online, and Field Experiments.” *Journal of Behavioral and Experimental Finance* 9: 88–97.

European Union. 2024. “Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024… (Artificial Intelligence Act), OJ L, 2024/1689, 12.7.2024.” Official Journal of the European Union.

Fehr, Ernst, and Klaus M. Schmidt. 1999. “A Theory of Fairness, Competition, and Cooperation.” *Quarterly Journal of Economics* 114(3): 817–868.

Fischbacher, Urs. 2007. “z-Tree: Zurich Toolbox for Ready-Made Economic Experiments.” *Experimental Economics* 10(2): 171–178.

Fish, Sara, Yannai A. Gonczarowski, and Ran I. Shorrer. 2024. “Algorithmic Collusion by Large Language Models.” arXiv:2404.00806.

Guan, Melody Y., Manas Joglekar, Eric Wallace, Saachi Jain, Boaz Barak, et al. 2024. “Deliberative Alignment: Reasoning Enables Safer Language Models.” arXiv:2412.16339.

Jagadeesh, Akshay V., Rahul K. Arora, Khaled Saab, Ali Malik, Mikhail Trofimov, Foivos Tsimpourlas, Johannes Heidecke, Karan Singhal. 2026. “Reinforcement Learning towards Broadly and Persistently Beneficial Models.” arXiv:2606.24014.

Lu, Wei, Amit Dhanda, Daniel L. Chen,and Christian B. Hansen. 2025. “Aligning Large Language Model Agents with Rational and Moral Preferences: A Supervised Fine-Tuning Approach.” arXiv:2507.20796.

Ouyang, Long, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, et al. 2022. “Training Language Models to Follow Instructions with Human Feedback.” *Advances in Neural Information Processing Systems* 35.

Parrish, Alicia, Angelica Chen, Nikita Nangia, et al. 2022. “BBQ: A Hand-Built Bias Benchmark for Question Answering.” Findings of the Association for Computational Linguistics: ACL 2022.

Röttger, Paul, Hannah Rose Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. 2024. “XSTest: A Test Suite for Identifying Exaggerated Safety Behaviours in Large Language Models.” NAACL 2024.

Souly, Alexandra, Qingyuan Lu, Dillon Bowen, et al. 2024. “A StrongREJECT for Empty Jailbreaks.” arXiv:2402.10260.

Tennant, Elizaveta, Stephen Hailes, and Mirco Musolesi. 2024. “Moral Alignment for LLM Agents.” arXiv:2410.01639.

U.S. Commodity Futures Trading Commission and U.S. Securities and Exchange Commission. 2010. “Findings Regarding the Market Events of May 6, 2010: Report of the Staffs of the CFTC and SEC to the Joint Advisory Committee on Emerging Regulatory Issues.”

van Leeuwen, Boris, and Ingela Alger. 2024. “Estimating Social Preferences and Kantian Morality in Strategic Interactions.” *Journal of Political Economy Microeconomics* 2(4): 665-706.

Wei, Jason, Nguyen Karina, Hyung Won Chung, et al. 2024. “Measuring Short-Form Factuality in Large Language Models.” arXiv:2411.04368.