

Test whom AI cooperation serves

Daniel L. Chen, Chris Hansen, Wei Lu

AI agents can cooperate with one another while harming society. Pricing algorithms can sustain prices above competitive levels without communicating¹.

Psychology and experimental economics offer tools for making alignment objectives explicit. Social-preference models specify how agents weigh their own returns against others' outcomes and generate training and evaluation data. Preliminary pricing simulations suggest that training with a Kantian objective—considering everyone following the same rule—lowers prices relative to the baseline. A preprint also reports gains from preference-based fine-tuning on benchmarks of bias, jailbreak resistance and overrefusal².

Independent evaluators should use controlled model access to estimate behavioral preferences and stress-test interactions across incentives, roles and unfamiliar settings. Publishing protocols and results would support replication while companies retain proprietary data and code.

Application-specific certification should be renewed after model updates. Passing these tests could support a rebuttable presumption of non-negligence, while mandatory insurance could price residual risk.

Human behavior provides benchmarks, not universal moral standards. Democratic institutions must determine acceptable trade-offs. Together, these steps offer a path to reproducible, legible, democratic AI governance grounded in social science theory and experiments.

References

1. Calvano, E., Calzolari, G., Denicolò, V. & Pastorello, S. Artificial Intelligence, Algorithmic Pricing, and Collusion. *American Economic Review* 110, 3267–3297 (2020). <https://pubs.aeaweb.org/doi/10.1257/aer.20190623>
2. Lu, W., Dhanda, A., Chen, D. L. & Hansen, C. B. Aligning Large Language Model Agents with Rational and Moral Preferences: A Supervised Fine-Tuning Approach. *arXiv:2507.20796v2* (2026). <https://arxiv.org/abs/2507.20796v2>