Bargaining Over Bias: Cooperative Contracts for AI Fairness

Manu, Aishika [1] **Batra, Pratishtha** [1]

Abstract

Fairness-critical AI systems require multi-stakeholder agreement on parameters such as disparity thresholds and harm weights, yet these are almost always set unilaterally by system designers. We formalise this as a mechanism design problem and introduce the Fairness Debt Contract (FDC): a negotiable tuple $\langle \delta, H, \alpha \rangle$ encoding a fairness threshold, harm weight, and accuracy budget. We compare four negotiation protocols and propose a hybrid Median-then-Nash (MtN) protocol. Agent-based simulations (150 rounds, 20 stakeholders) show that MtN produces approximately 60% higher (1.6×) social welfare (defined here as the sum of stakeholder preference satisfaction at the negotiating table, distinct from downstream demographic outcomes) than Nash bargaining (4.06 vs. 2.54) and converges in every round, whereas unanimity never converges. MtN's first stage, median scoping, is verifiable and auditable, making it a candidate governance primitive for Technical AI Governance (verifiability, operationalisation) and Cooperative AI (bargaining, trust).

1. Introduction

Consequential AI systems embed fairness parameters that are value-laden policy choices. Who should decide that the acceptable disparity between demographic groups is $\delta = 0.10$ rather than 0.20? In practice, these choices are delegated to engineers, accumulating *fairness debt* (de Souza Santos et al., 2025). Affected communities, regulators, platform operators, and advocacy groups each hold legitimate but conflicting interests over these parameters.

Existing work addresses fairness as an optimisation problem (Chouldechova, 2017; Yang et al., 2024) but treats the parameters themselves as fixed. We argue that setting these parameters is the governance problem. This reframing connects to a broader tradition: participatory machine learning (Birhane et al., 2022; Sloane et al., 2022) and value-sensitive design (Friedman et al., 2019) have long argued that affected communities must have standing in system design; our contribution is to formalise this participation as an auditable, manipulation-bounded mechanism.

In mechanism design, hybrid protocols that pair a strategy-proof scoping stage with an efficient selection stage have been studied in the context of committee elections and resource allocation (Manjunath, 2012; Sui et al., 2013). To our knowledge, this structure has not been applied to AI fairness parameter negotiation, nor framed as a governance artefact.

Contributions. We define the Fairness Debt Contract (FDC), a three-tuple $\langle \delta, H, \alpha \rangle$ that makes the accuracy–fairness trade-off a negotiable object. We propose **Median-then-Nash (MtN)**, a two-stage hybrid that achieves strategy-proof scoping followed by Pareto-optimal selection. Agent-based simulations confirm MtN achieves approximately 60% higher stakeholder preference satisfaction than Nash while dominating all protocols on convergence. We discuss the limitations of these simulation results and the substantial additional work required before deployment.

2. Fairness Debt Contract

Let $\Delta(t)$ denote the demographic disparity of an AI system at time t. Following de Souza Santos et al. (2025), we define fairness debt as the integral of harm exceeding a threshold:

$$D(T) = \int_0^T H \cdot \max(0,\, \Delta(t) - \delta)\, dt$$

where $\delta \in [0.02, 0.30]$ is the disparity threshold, $H \in [0.1, 0.9]$ is the harm weight, and $\alpha \in [0.01, 0.10]$ is the *accuracy budget*. The **Fairness Debt Contract** is the negotiable tuple:

$$\text{FDC} = \langle \delta,\, H,\, \alpha \rangle.$$

Each stakeholder i holds a private ideal point $x_i^* = (\delta_i^*, H_i^*, \alpha_i^*)$. The *disagreement point* is the status quo: $\delta_0 = 0.20$, $H_0 = 0.1$, $\alpha_0 = 0.01$.

[1]Indira Gandhi Delhi Technical University for Women, Delhi, India. Correspondence to: Aishika Manu <aishika096btmae23@igdtuw.ac.in>.

Second Workshop on Technical AI Governance Research (TAIGR) @ ICML 2026, Seoul, South Korea. 2026. Copyright 2026 by the author(s).

Table 1. Protocol comparison.

Protocol	Pareto efficient	Manipulation resistance	Convergence
Unanimity	–	Yes (veto)	No
Majority	Partial	No	Yes
Iterated median	Partial	Yes (Moulin, 1980)	Yes
Nash bargaining	Yes	No	Yes
MtN (ours)	**Yes**	**Bounded (Prop. 1)**	**Yes**

We model four stakeholder types calibrated to content-moderation deployment: *platform operator* (moderate δ, balanced), *civil-society advocate* (strict δ, high H), *regulator* (moderate δ, maximal H for compliance with instruments such as the EU DSA), and *creator/user community* (permissive δ, accuracy-preserving).

A note on "social welfare." Throughout this paper, social welfare refers to the sum of preference satisfaction of the stakeholders at the negotiating table – the standard mechanism design usage. It does not directly measure outcomes for the demographic groups whose treatment δ governs. The advocate archetype is intended to proxy for those interests, but it remains a synthetic construct. Improving the connection between negotiated parameter values and downstream demographic outcomes is an important direction for future work.

3. Negotiation Protocols

We evaluate five protocols (Table 1). Unanimity requires all stakeholders to accept. Majority voting applies per-dimension majority. Iterated median selects the coordinate-wise median of reported ideals. Nash bargaining selects $\arg\max \prod_i g_i(c)$ (Nash, 1950).

Median-then-Nash (MtN). Stage 1 computes the coordinate-wise median m of reported ideals; Stage 2 applies Nash bargaining within the bandwidth ball.

Relationship to iterated median. Both MtN Stage 1 and the iterated median are strategy-proof by Moulin (1980) and converge in 100% of rounds. The key distinction is efficiency: the iterated median is only partially Pareto-efficient, whereas MtN's Nash Stage 2 selects from the Pareto frontier within the scoped region. A direct welfare comparison between MtN and iterated median is not reported here because social welfare is not a primary metric for voting-style baselines in our simulation design; we identify this as a gap and a priority for follow-on work.

4. Proposition 1

Let agents have single-peaked preferences on each coordinate, with truthful ideals drawn i.i.d. from a distribution with bounded density on the valid range. Let m be the coordinate-wise median of n truthful reports, and let m' be the median when a coalition of $k < n/2$ agents misreports arbitrarily. Let c^* and c' denote the MtN outcomes under truthful and manipulated reports respectively. Then

$$\|c' - c^*\|_\infty \;\le\; \|m' - m\|_\infty + 2b \;=\; O(k/n) + O(b),$$

where b is the Stage 2 bandwidth.

Proof sketch. *Stage 1.* The coordinate-wise median is strategy-proof for single-peaked preferences (Moulin, 1980). A coalition of size $k < n/2$ cannot move the median beyond the $(n/2 - k)$-th and $(n/2 + k)$-th truthful order statistics. Under bounded density, adjacent order statistics near the median are separated by $O(1/n)$, giving $\|m' - m\|_\infty = O(k/n)$. *Stage 2.* By construction, MtN selects from $B(m', b)$, so $\|c' - m'\|_\infty \le b$ and $\|c^* - m\|_\infty \le b$. The triangle inequality yields the claim.

Limitations of the bound. Proposition 1 assumes separable, single-peaked utilities on each coordinate. This excludes interaction effects between parameters – for example, a stakeholder who will only accept higher δ if H also increases. Such coupled preferences would require a richer utility model (e.g., supermodular or lexicographic) and a correspondingly different manipulation analysis. We flag this as the primary theoretical gap.

Regarding collusion: Proposition 1 bounds manipulation for any coalition of size $k < n/2$. For larger coordinated coalitions ($k \ge n/2$), the bound no longer holds and median capture becomes possible, as in any median-based mechanism. In our simulation the misreporting fraction is fixed at 25% ($k = 5/20$), well within the safe regime. Robustness to larger or more sophisticated coalitions is an open problem.

Regarding adaptive bandwidth: making b a function of reported dispersion could improve welfare at the cost of weakening the $O(b)$ component of the manipulation bound. The tradeoff tips when the marginal welfare gain from enlarging b is offset by the increased manipulation surface; we have not characterised the optimal b analytically, and we treat this as future work.

5. Experimental Setup

Agents. 20 stakeholders (5 per archetype). Each agent's ideal is drawn from a Gaussian centred on the archetype ideal with $\sigma = 0.35$, clipped to valid ranges. 300 contracts sampled uniformly at random per round: $\delta \in [0.02, 0.30]$, $H \in [0.1, 0.9]$, $\alpha \in [0.01, 0.10]$.

Strategic manipulation. In the strategic condition, 25% of agents misreport their ideal. Platform operators inflate δ by a factor of $1.5\times$; advocates deflate δ by $1.5\times$. **Welfare is evaluated against true ideals.**

Table 2. Simulation results across 150 rounds. Pareto distance = 0 by construction of the sampled frontier. †Welfare not primary metric for voting baselines.

Protocol	Social welfare ↑	Convergence
Unanimity	—	0.0%
Majority	—†	100%
Iterated median	—†	100%
Nash (truthful)	2.540 ± 1.256	100%
MtN (truthful)	$\mathbf{4.056 \pm 1.396}$	100%
Nash (strategic)	2.516 ± 1.259	100%
MtN (strategic)	$\mathbf{4.056 \pm 1.396}$	100%

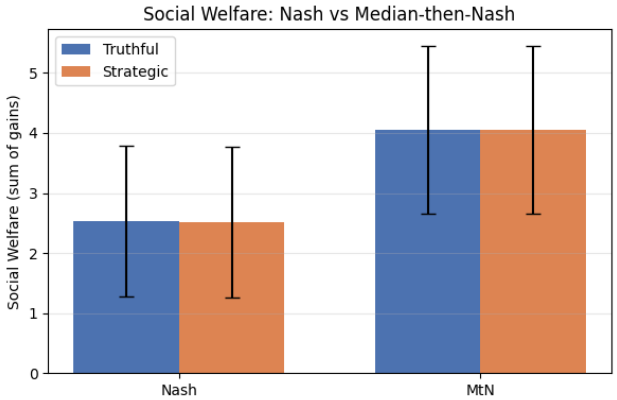

Figure 1. Social welfare comparison (truthful and strategic conditions) for Nash and MtN. Error bars show ± 1 standard deviation across 150 rounds. MtN's stability under the strategic condition is consistent with the $O(b/n)$ manipulation bound derived in Proposition 1.

Metrics. Pareto distance (gain-space Euclidean distance to nearest efficient contract), social welfare (sum of true gains), convergence rate (fraction of valid outcomes).

Rounds. 150 independent rounds; agents re-sampled each round.

Scope of inference. The simulation is designed to test whether MtN's mathematical properties (bounded manipulation, Pareto efficiency within scope, convergence) hold under the assumed utility model. The stakeholder archetypes, their Gaussian ideal distributions, and the noise parameters are all researcher-specified. Results should therefore be interpreted as confirming internal consistency under these assumptions, not as evidence of real-world performance. The 1.6× welfare figure is a property of this simulation family; ecological validity with actual political dynamics, institutional constraints, and heterogeneous stakeholder power remains to be established.

6. Results

Convergence. Unanimity never converges (0%); all others converge in 100% of rounds (Table 2).

Social welfare. MtN achieves a mean welfare of 4.056 ± 1.396 vs. Nash's 2.540 ± 1.256 – an approximately 60% increase (1.6×, $p < 0.001$, paired t-test). MtN's median scoping prevents selection of Pareto contracts that disproportionately benefit narrow coalitions. Under strategic manipulation, MtN's welfare is unchanged (4.056 ± 1.396), while Nash's is minimally affected (2.516 ± 1.259), consistent with the $O(b/n)$ bound.

7. Governance and Cooperative AI Implications

Technical AI Governance. Stage 1 admits a standard commit–reveal protocol and the median can be logged immutably. This addresses *verification* and *operationalisation* as challenges central to technical governance (Lim & Lim, 2024). From a cooperative bargaining perspective (Dafoe et al., 2020; Conitzer et al., 2024), MtN occupies a position between the median voter solution (strategy-proof but not Pareto-optimal) and Nash bargaining (Pareto-optimal but manipulable), achieving both within the scoped region. Median scoping further protects stakeholders with less bargaining power – e.g., advocates in platform-dominated settings – by anchoring selection near the social median, aligning with participatory frameworks (Vineis & Perelli, 2024; Birhane et al., 2022; Sloane et al., 2022).

Situating MtN in the mechanism design literature. Hybrid mechanisms combining a strategy-proof scoping stage with an efficient selection stage have precedent in social choice (Manjunath, 2012; Sui et al., 2013). MtN's specific contribution is the application of this structure to multi-dimensional fairness parameter negotiation and its framing as an auditable governance artefact. A more thorough survey of related constructions – particularly in participatory ML (Birhane et al., 2022) and value-sensitive design (Friedman et al., 2019) – is warranted and will be developed in the full version of this work.

8. Limitations and Future Work

Utility model. Separable single-peaked utilities exclude interaction effects between δ, H, and α – e.g., stakeholders who tolerate higher disparity only when harm weight rises. Extending MtN to supermodular or lexicographic preferences, and re-deriving Proposition 1 under these richer models, is the primary theoretical priority.

Fixed bandwidth. The bandwidth b = 0.08 is fixed throughout. Adaptive b could improve welfare but weakens the manipulation bound. Characterising the optimal b remains an open question.

Synthetic stakeholders. Our stakeholders are synthetic archetypes drawn from researcher-specified Gaussians. The simulation confirms internal mathematical consistency; it

does not test real political dynamics, institutional power asymmetries, or the effects of legal standing on participation. A pilot deployment with an actual platform, civil-society advocate, and regulator is the necessary stress test. Until such a pilot exists, the welfare numbers should be treated as theoretical rather than empirical benchmarks.

Connection to demographic outcomes. As noted in Section 2, the social welfare metric measures preference satisfaction among negotiating parties, not outcomes for the demographic groups governed by δ. Developing metrics that connect the negotiated FDC directly to downstream harm reduction is essential for the governance case.

9. Conclusion

Fairness parameters decide who gets misclassified, flagged, or screened out. Today those choices are made unilaterally. MtN makes them negotiable: affected communities, advocates, and regulators participate in setting the parameters that govern them, with manipulation bounded by $O(b) + O(k/n)$ and approximately $1.6\times$ the preference satisfaction of pure Nash bargaining under our simulation assumptions. The negotiation itself becomes an audit artefact. We view MtN as a useful first step that requires real-stakeholder validation before deployment.

Impact Statement

Fairness parameters in content moderation, hiring, and risk assessment determine who is protected and who is harmed, yet are set without input from affected communities. MtN distributes this power via a verifiable, negotiable mechanism. It is a technical primitive, not a substitute for democratic process. Risks include capture of the bandwidth parameter by well-resourced stakeholders, exclusion of communities without formal standing, and the gap between stakeholder preference satisfaction and outcomes for the demographic groups the parameters govern.

LLM Usage Statement

We used Claude Opus 4.5 to assist with writing and editing prose for clarity and concision, and to debug LaTeX compilation issues. All formatting was done independently by the authors. All research ideas, the Fairness Debt Contract formulation, the Median-then-Nash protocol design, the theoretical analysis in Proposition 1, and the experimental design were conceived and carried out by the authors, who take full responsibility for all content in this paper.

References

Birhane, A., Isaac, W., Prabhakaran, V., et al. The values encoded in machine learning research. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2022.

Chouldechova, A. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 5(2):153–163, 2017.

Conitzer, V., Ge, L., Minsky, M., et al. Social choice should guide ai alignment in dealing with diverse human feedback. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, Vienna, Austria, 2024. PMLR.

Dafoe, A., Hughes, E., Bachrach, Y., et al. Open problems in cooperative ai. In *NeurIPS Workshop on Cooperative AI*, Vancouver, Canada, 2020.

de Souza Santos, R. et al. Software fairness debt: Concepts, measurement, and management. *ACM Transactions on Software Engineering and Methodology*, 34(3): 1–35, 2025.

Friedman, B., Hendry, D. G., and Borning, A. A survey of value sensitive design methods. *Foundations and Trends in Human–Computer Interaction*, 11(2-3):63–125, 2019.

Lim, E. C. N. and Lim, C. E. D. Polycentric ai governance: A multi-stakeholder approach to distributed responsibility and ethical technology management. *International Journal of Advanced AI Applications*, 2024.

Manjunath, V. Efficient and strategy-proof social choice when preferences are single-dipped. *International Journal of Game Theory*, 41(3):595–617, 2012.

Moulin, H. On strategy-proofness and single peakedness. *Public Choice*, 35(4):437–455, 1980.

Nash, J. F. The bargaining problem. *Econometrica*, 18(2): 155–162, 1950.

Sloane, M., Moss, E., Awomolo, O., and Forlano, L. Participation is not a design fix for machine learning. In *Proceedings of the ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO)*, 2022.

Sui, X., Boutilier, C., and Sandholm, T. Analysis and optimization of multi-dimensional percentile mechanisms. In *Proceedings of the 23rd International Joint Conference on Artificial Intelligence (IJCAI)*, 2013.

Vineis, V. and Perelli, G. Beyond predictions: A participatory framework for multi-stakeholder decision-making. arXiv preprint, 2024.

Yang, X. et al. Fairness-aware meta-learning via nash bargaining. In *Advances in Neural Information Processing Systems (NeurIPS)*, San Diego, CA, 2024. Curran Associates.