Supporting Information for

A Computational Theory of Cooperation under Moral Uncertainty

Avi Sharma, Dasari Sai Harsh, Jainendra Shukla

Corresponding Author name.
E-mail: jainendra@iiitd.ac.in

This PDF file includes:

Supporting text
Figs. S1 to S6
Legend for Dataset S1

Other supporting materials for this manuscript include the following:

Dataset S1

Supporting Information Text

S1. Model Details

This section develops the formal components of the Trust-Gated Moral Inference (TGMI) model. We define the game environment, fairness mappings, belief representations, trust dynamics, and the virtual bargaining mechanism. These definitions correspond to Eqs. (1)–(4) in the main text and underlie Algorithm 1.

S1.1 Game Structure. Each interaction is a two-player continuous-action game

$$G = (\mathcal{A}_i,\, \mathcal{A}_j,\, R_i,\, R_j,\, \mathcal{F}),$$

where, for each agent i:

- $\mathcal{A}_i \subseteq [0,1]$ is the action space,
- $R_i(a_i, a_j) \in [0,1]$ is a normalized payoff function,
- $\mathcal{F}$ is the set of fairness principles.

A joint action is $a = (a_i, a_j)$. Action errors occur with probability ε_a; the realized action is

$$\tilde{a}_i \sim \mathcal{N}(a_i, \varepsilon_a) \quad \text{projected to } [0,1].$$

S1.2 Moral Priors and Fairness Mappings. Each agent i has an intrinsic moral prior $B_i(\phi)$ over $\phi \in \mathcal{F}$, initialized from a symmetric Dirichlet distribution.

Fairness functions satisfy:

$$0 \le F_\phi(a_i, a_j) \le 1.$$

In the MGG:

$$F_{\text{Max-Sum}} = \frac{R_i + R_j}{R_{\max}}, \qquad F_{\text{Equal-Split}} = 1 - \frac{|R_i - R_j|}{R_{\max}}, \qquad F_{\text{Rawls}} = \frac{\min(R_i, R_j)}{R_{\max}}.$$

S1.3 Beliefs About Others' Moral Priors. Agent i maintains a belief distribution $\hat{B}_{i\to j}(\phi)$. Initial beliefs are

$$\hat{B}_{i\to j}(\phi) \sim \text{Dirichlet}(1,\ldots,1), \qquad \mathbb{E}[\hat{B}_{i\to j}(\phi)] = 1/|\mathcal{F}|.$$

Confidence is

$$c_i = 1 - \frac{H(\hat{B}_{i\to j})}{\log|\mathcal{F}|}.$$

S1.4 Trust and Effective Cooperation Weight.

$$\tau_i^{(t+1)} = (1-\eta)\tau_i^{(t)} + \eta\, s_i^{(t)}, \qquad s_i^{(t)} = \exp(-\lambda_{\text{dev}} d_i^{(t)}).$$

The cooperation weight is

$$\kappa_i = \tau_i\, c_i.$$

S1.5 Trust-Gated Moral Utility.

$$U_i(a_i, a_j) = \sum_{\phi\in\mathcal{F}} \Big[(1-\kappa_i)B_i(\phi) + \kappa_i\, \hat{B}_{i\to j}(\phi)\Big] F_\phi(a_i, a_j).$$

$$\widehat{U}_j^{(i)}(a_j, a_i) = \sum_{\phi\in\mathcal{F}} \hat{B}_{i\to j}(\phi)\, F_\phi(a_j, a_i).$$

S1.6 Fairness Deviation.

$$U_i^F(a_i, a_j) = \sum_{\phi\in\mathcal{F}} B_i(\phi) F_\phi(a_i, a_j).$$

$$d_i^{(t)} = \max_{a'_j} U_i^F(a_i^{\text{VB}}, a'_j) - U_i^F(a_i^{\text{VB}}, a_j^{\text{VB}}).$$

S1.7 CK-ToM Belief Update.

$$\tilde{B}_{i\to j}^{(t+1)}(\phi) = \hat{B}_{i\to j}^{(t)}(\phi)\, \exp\!\big(\beta\, \alpha\tau_i^{(t)} F_\phi(a_i^{\text{VB}}, a_j^{\text{VB}})\big)\, [B_i(\phi)]^{1-\alpha\tau_i^{(t)}}.$$

$$\hat{B}_{i\to j}^{(t+1)}(\phi) = \frac{\tilde{B}_{i\to j}^{(t+1)}(\phi)}{\sum_{\phi'\in\mathcal{F}} \tilde{B}_{i\to j}^{(t+1)}(\phi')}.$$

S1.8 Virtual Bargaining Equilibrium (VBE).

$$(a_i^{\mathrm{VB}}, a_j^{\mathrm{VB}}) = \arg\max_{a_i,a_j} [(U_i(a_i,a_j) - d_i)_+]^{\gamma} [(\widehat{U}_j^{(i)}(a_j,a_i) - d_j)_+]^{1-\gamma}.$$

$$d_i^{(t+1)} = U_i^F(a_i^{\mathrm{VB}}, a_j^{\mathrm{VB}}).$$

S1.9 Logging, Errors, and Observability.

$$\{\tau_i^{(t)}, c_i^{(t)}, \kappa_i^{(t)}, d_i^{(t)}, U_i, U_i^F, R_i, \hat{B}_{i\to j}^{(t)}\}.$$

Action error ε_a and perception error ε_p apply independently to actions and fairness signals.

S2. Theoretical Properties of Trust-Gated Moral Inference

S2.1 Preliminaries and Assumptions. We analyze the learning and decision dynamics of the Trust-Gated Moral Inference (TGMI) model defined in Algorithm 1. Agents interact in a continuous-action game with joint action space

$$\mathcal{A} = \mathcal{A}_i \times \mathcal{A}_j,$$

and a finite set of fairness principles $\mathcal{F}$.

Each agent i evaluates joint actions through a trust-gated moral utility

$$U_i(a_i,a_j) = \sum_{\phi\in\mathcal{F}} \Big[(1-\kappa_i)\, B_i(\phi) + \kappa_i\, \hat{B}_{i\to j}(\phi)\Big] F_\phi(a_i,a_j),$$

where $\kappa_i = \tau_i c_i \in [0,1]$ combines trust τ_i and confidence c_i.

Agent i's internal model of partner j's utility is given by

$$\widehat{U}_j^{(i)}(a_j,a_i) = \sum_{\phi\in\mathcal{F}} \hat{B}_{i\to j}(\phi)\, F_\phi(a_j,a_i),$$

which represents i's best estimate of how j evaluates outcomes under moral uncertainty.

We impose the following assumptions.

(A1) Compactness. The action spaces $\mathcal{A}_i$ and $\mathcal{A}_j$ are compact subsets of $\mathbb{R}^d$.

(A2) Continuity. Each fairness function $F_\phi(a_i,a_j)$ is continuous in (a_i,a_j). By construction, both U_i and $\widehat{U}_j^{(i)}$ inherit continuity from F_ϕ and the belief distribution $\hat{B}_{i\to j}$.

(A3) Boundedness. All fairness utilities satisfy $F_\phi(a_i,a_j) \in [0,1]$. Reservation utilities $d_i^{(t)}$ are initialized at zero and updated via realized fairness values, hence remain bounded in $[0,1]$.

These assumptions are satisfied by the Moral Game Generator used in all simulations.

S2.2 Existence of a Virtual Bargaining Equilibrium. At each interaction, agent i computes a counterfactual joint action via virtual bargaining:

$$(a_i^{\mathrm{VB}}, a_j^{\mathrm{VB}}) = \arg\max_{a_i,a_j} \Big[(U_i(a_i,a_j) - d_i)_+\Big]^{\gamma} \Big[(\widehat{U}_j^{(i)}(a_j,a_i) - d_j)_+\Big]^{1-\gamma},$$

where $(x)_+ = \max(x,0)$ and $\gamma \in (0,1)$.

Proposition S2.1 (Existence of Virtual Bargaining Equilibrium). Under assumptions (A1)–(A3), there exists at least one Virtual Bargaining Equilibrium (VBE).

Proof. The Nash-style product objective is continuous over the compact domain $\mathcal{A}_i \times \mathcal{A}_j$. By the Weierstrass extreme value theorem, a maximizer exists. □

This result guarantees that TGMI's coordination step is well-defined at every round. We do not claim uniqueness or Nash equilibrium refinements.

S2.3 Bounded Loss and Gain under Unilateral Deviation. We next examine the consequences of unilateral deviation from the virtual bargaining outcome.

Lemma S2.2 (Bounded Loss and Gain). For any agent i, the maximum possible gain or loss in trust-gated moral utility from a unilateral deviation is bounded by a finite constant independent of the partner's strategy.

Proof. By boundedness of fairness utilities and belief weights,

$$0 \le U_i(a_i, a_j) \le 1 \quad \text{for all } (a_i, a_j).$$

Hence, for any deviation $a_i' \in \mathcal{A}_i$,

$$\left| U_i(a_i', a_j) - U_i(a_i^{\mathrm{VB}}, a_j^{\mathrm{VB}}) \right| \le 1.$$

Thus, no single deviation can produce arbitrarily large gains or losses. □

Interpretation. This bound is absolute rather than asymptotic: it guarantees that unilateral deviations cannot cause unbounded exploitation or catastrophic losses, without invoking equilibrium refinements or regret notions.

S2.4 Joint Welfare and Coordination Guarantee. Although utilities are bounded below by construction, the virtual bargaining mechanism provides a stronger coordination property.

Proposition S2.3 (Fairness-Based Coordination). Whenever a joint action exists that strictly improves both agents' fairness utilities above their reservation levels, virtual bargaining selects a joint action that is not Pareto-dominated in fairness space.

Proof. The bargaining objective maximizes a product of gains above reservation utilities. Any outcome that is Pareto-dominated in fairness space yields a strictly smaller product than a jointly improving alternative, and therefore cannot be selected. □

Corollary. Virtual bargaining excludes mutually destructive outcomes whenever any fairness-consistent improvement is feasible.

This result establishes that TGMI promotes coordination rather than merely avoiding negative outcomes.

S2.5 Convergence of Trust–Belief Dynamics. Belief updating in TGMI follows a CK-ToM multiplicative update rule:

$$\hat{B}_{i\to j}^{(t+1)}(\phi) \propto \hat{B}_{i\to j}^{(t)}(\phi)\, \exp\!\big(\beta\, \alpha \tau_i^{(t)} F_\phi(a^{\mathrm{VB}})\big)\, [B_i(\phi)]^{1-\alpha\tau_i^{(t)}}.$$

Proposition S2.4 (Belief Stabilization under Trust Convergence). If trust converges to a limit $\tau_i^{(t)} \to \tau_i^\star$, then the belief sequence $\hat{B}_{i\to j}^{(t)}$ converges to a stationary distribution minimizing Kullback–Leibler divergence to a weighted combination of observed fairness evidence and the intrinsic prior.

Proof Sketch. The update is equivalent to a multiplicative-weights procedure with a stationary learning rate. Standard results imply convergence to a fixed point minimizing a convex KL objective. □

The assumption $\tau_i^{(t)} \to \tau_i^\star$ holds whenever fairness deviations stabilize, a condition observed empirically across all stationary regimes.

S3. Simulation Methods and Implementation Details

This appendix describes the simulation procedures used to evaluate TGMI in the Moral Game Generator (MGG), including: (i) intragenerational learning dynamics, (ii) intergenerational evolutionary simulations, (iii) parameter sweeps and robustness checks, and (iv) implementation details ensuring reproducibility. The aim is to specify modeling components precisely while keeping the computational framework transparent.

S3.1 Action Spaces and Numerical Resolution. The continuous action space is discretized as

$$\mathcal{A}_i = \{0,\, \Delta,\, 2\Delta, \ldots, 1\}, \qquad \Delta = 0.05.$$

This resolution captures the payoff and fairness landscapes while ensuring tractable Virtual Bargaining (VB) optimization. Robustness checks in Section S5 confirm stable outcomes for $\Delta = 0.025$.

All fairness functions $F_\phi(a_i, a_j)$ and payoffs $R_i(a_i, a_j)$ are evaluated on this grid; interpolation is unnecessary because VB optimization searches only over the discrete space.

S3.2 Sampling Games from the Moral Game Generator (MGG). Each sampled interaction G consists of four components.

(1) Payoff surface. A normalized payoff function

$$R_i(a_i, a_j) \in [0, 1]$$

is drawn from one of four archetypes: **Dilemma**, **Assurance**, **Bargain**, and **Public-Goods**. Each surface is sampled from a parametric family over $[0,1]^2$ with parameters drawn uniformly.

(2) Fairness environment. The fairness basis is

$$\mathcal{F} = \{\text{Max-Sum, Equal-Split, Rawls}\},$$

with mappings

$$F_{\mathrm{MaxSum}} = \frac{R_i + R_j}{R_{\max}}, \qquad F_{\mathrm{EqSplit}} = 1 - \frac{|R_i - R_j|}{R_{\max}}, \qquad F_{\mathrm{Rawls}} = \frac{\min(R_i, R_j)}{R_{\max}}.$$

(3) Moral priors. Each agent receives

$$B_i \sim \text{Dirichlet}(1, 1, 1),$$

yielding heterogeneous but unbiased fairness weights.

(4) Noise.

- **Action error** ϵ_a: with probability ϵ_a, the selected action is replaced by a uniformly random action.
- **Perception noise** ϵ_p: fairness values are perturbed as

$$\widetilde{F}_\phi = (1 - \epsilon_p)F_\phi + \epsilon_p \cdot U(0, 1).$$

Each round samples a new payoff–fairness environment, following the Bayesian game-generator logic of (**?**).

S3.3 Intragenerational Learning Simulations. Two agents i and j interact for T rounds. Each round proceeds as follows:

1. **Utility computation.**
$$U_i(a_i, a_j), \qquad \widehat{U}_j^{(i)}(a_j, a_i).$$
2. **Virtual Bargaining (VB).**
$$(a_i^{\text{VB}}, a_j^{\text{VB}}) = \arg\max_{a_i, a_j} \left[(U_i - d_i)_+\right]^\gamma \left[(\widehat{U}_j^{(i)} - d_j)_+\right]^{1-\gamma}.$$
3. **Action realization with noise.** With probability $1 - \epsilon_a$, each agent executes its VB action; otherwise a uniformly random action is substituted.
4. **Fairness deviation.**
$$d_i^{(t)} = \max_{a'_j} U_i^F(a_i^{\text{VB}}, a'_j) - U_i^F(a_i^{\text{VB}}, a_j^{\text{real}}).$$
5. **Trust update.**
$$\tau_i^{(t+1)} = (1 - \eta)\tau_i^{(t)} + \eta \, \exp(-\lambda_{\text{dev}} d_i^{(t)}).$$
6. **Belief update (CK-ToM).**
$$\tilde{B}_{i \to j}^{(t+1)}(\phi) = \hat{B}_{i \to j}^{(t)}(\phi) \, \exp(\beta\alpha\tau_i^{(t)} F_\phi) \, [B_i(\phi)]^{1-\alpha\tau_i^{(t)}},$$
normalized as
$$\hat{B}_{i \to j}^{(t+1)}(\phi) = \frac{\tilde{B}_{i \to j}^{(t+1)}(\phi)}{\sum_{\phi' \in \mathcal{F}} \tilde{B}_{i \to j}^{(t+1)}(\phi')}.$$
7. **Confidence and cooperation weight.**
$$c_i^{(t+1)} = 1 - \frac{H(\hat{B}_{i \to j}^{(t+1)})}{\log |\mathcal{F}|}, \qquad \kappa_i^{(t+1)} = \tau_i^{(t+1)} c_i^{(t+1)}.$$
8. **Reservation utility update.**
$$d_i^{(t+1)} = U_i^F(a_i^{\text{VB}}, a_j^{\text{VB}}).$$

Each condition is replicated for 1,000 random seeds, logging:

$$\{\tau_i^{(t)}, c_i^{(t)}, \kappa_i^{(t)}, d_i^{(t)}, U_i, U_i^F, R_i, \hat{B}_{i \to j}^{(t)}\}.$$

S3.4 Baseline and Ablation Agents. We compare TGMI against standard reciprocity strategies:

$$\text{TFT, GTFT, WSLS, Forgiver, AllC, AllD, ZD-Extort,}$$

following (**? ?**).

TGMI ablations isolate key components:

1. **No-trust:** τ_i fixed at τ_0.
2. **No-belief:** $\hat{B}_{i \to j}$ fixed at the Dirichlet prior.
3. **No-bargaining:** VB replaced with unilateral maximization of $U_i(a_i, a_j)$.

S3.5 Evolutionary Simulations (Moran Process). We use a finite-population Moran process with

$$N = 10, \qquad \mu = 10^{-3}, \qquad s = 2.$$

Each generation includes:

1. Interaction of all agents for T rounds in independently sampled MGG games.
2. Computation of fairness-weighted returns:

$$\Phi_i = \frac{1}{T}\sum_{t=1}^{T}\Big[(1-\omega)R_i^{(t)} + \omega U_i^{F,t}\Big].$$

Softmax selection.

$$\pi_k(\mathbf{n}) = \frac{\exp(s\,\bar{\Phi}_k(\mathbf{n}))}{\sum_m \exp(s\,\bar{\Phi}_m(\mathbf{n}))}.$$

Agents adopt type k with probability $(1-\mu)\pi_k$ or mutate uniformly with probability μ.

Stationary distribution. The stationary distribution x^* satisfies

$$x^* = x^* P,$$

computed via power iteration (tolerance 10^{-9}). The abundance of type k is

$$\bar{x}_k^* = \sum_{\mathbf{n}} \frac{n_k}{N}\, x^*(\mathbf{n}).$$

S3.6 Parameter Settings. Default parameters (unless varied):

Parameter	Meaning	Default
β	Evidence sensitivity	5
α	Trust-based anchoring	0.5
η	Trust update rate	0.1
λ_{dev}	Deviation penalty	8
γ	VB asymmetry	0.5
τ_0	Initial trust	0.5
ϵ_a	Action noise	0.05
ϵ_p	Perception noise	0.00
T	Rounds per generation	20

S3.7 Software and Reproducibility. Simulations were implemented in Python using NumPy and JAX backends. All experiments use fixed random seeds. The codebase includes modules for:

- MGG generation,
- TGMI agent implementation,
- VB solver,
- evolutionary engine,
- figure-generation scripts.

All scripts and configuration files will be released upon publication.

S4. Supplementary Figures and Extended Simulation Results

This section provides additional simulation analyses complementing the main-text results. Figures S4.1–S4.6 illustrate (i) intragenerational learning dynamics, (ii) virtual bargaining behavior across payoff and fairness gradients, (iii) robustness to noise and perceptual uncertainty, (iv) evolutionary stability under the Moral Game Generator (MGG), and (v) the effects of increasing moral heterogeneity in the population. All simulations use the TGMI algorithm defined in Algorithm 1, unless otherwise specified. Parameters, seeds, and full configuration files are available in the project repository.

S4.1. Intragenerational Trust and Belief Dynamics. In this section we illustrate how a single TGMI agent updates its belief distribution $\hat{B}_{i\to j}$ and trust variable τ_i over repeated interactions with partners of varying moral types. Each trajectory reports the mean of 1,000 Monte Carlo runs with 95% confidence intervals.

Fig. S1. Fig. S4.1. Intragenerational learning dynamics. (A) Belief updates when interacting with a fairness-aligned partner. (B) Trust and belief collapse under exploitation by a self-interested partner. (C) Partial convergence when interacting with a morally dissimilar partner. Across all cases, TGMI converges toward accurate moral inference while adapting trust in a directionally appropriate manner.

S4.2. Virtual Bargaining Landscapes. To visualize the counterfactual agreement implied by TGMI, we compute the full virtual bargaining landscape $V(a_i, a_j)$ for sampled payoff–fairness pairs. These heatmaps illustrate how TGMI selects joint actions that maximize trust-gated moral utility rather than unilateral payoff or unilateral fairness.

Fig. S2. Fig. S4.2. Virtual bargaining landscapes. Heatmaps show the bargaining objective $\left[(U_i - d_i)_+\right]^{\gamma}\left[(\widehat{U}_j^{(i)} - d_j)_+\right]^{1-\gamma}$ for three representative games: Dilemma, Bargain, and Rawlsian conflict. Colored dots indicate selected $(a_i^{\mathrm{VB}}, a_j^{\mathrm{VB}})$. TGMI systematically avoids exploitative action profiles and selects morally endorsable compromises even in asymmetric games.

S4.3. Robustness to Noise and Perceptual Uncertainty. We test TGMI under (i) action error ε_a and (ii) fairness-perception noise ε_p, as defined in Section S2. The key robustness measure is whether trust and cooperation remain positive over long horizons.

Fig. S3. Fig. S4.3. Noise robustness. (A) Trust trajectories as a function of action error ε_a. (B) Cooperation frequency under fairness-perception noise ε_p. (C) Combined noise: $(\varepsilon_a, \varepsilon_p)$ grid showing stability region where cooperative VBE remains incentive-compatible. TGMI tolerates substantially more noise than reciprocity-based baselines because trust updates depend on fairness deviation rather than immediate payoff.

S4.4. Evolutionary Dynamics in the Moral Game Generator. Using the Moran process described in the main text and Appendix S2, we evaluate the steady-state abundance of TGMI and baseline agents across varying interaction length T, noise, and observability. Results below are averaged over 200 replicates per population composition.

Fig. S4. Fig. S4.4. Evolutionary abundance under the MGG. (A) TGMI steady-state abundance as a function of game length T. (B) Evolutionary stability under action error. (C) Stability map over (T, ε_a) with regions of TGMI dominance (shaded). TGMI outcompetes payoff-based reciprocal strategies except under extreme noise.

S4.5. Effects of Moral Heterogeneity. We evaluate the impact of increasing the variance of moral priors B_i (by adjusting the Dirichlet concentration parameter α_{Dir}). As heterogeneity increases, partners' fairness norms diverge more strongly, providing a stress test for moral inference.

Fig. S5. Fig. S4.5. Impact of moral heterogeneity. (A) Distribution of sampled moral priors for increasing Dirichlet variance. (B) Cooperation rates across heterogeneity levels. (C) Trust–belief stability diagram showing that TGMI maintains positive cooperation for moderate heterogeneity, with graceful degradation thereafter. This demonstrates TGMI's ability to synchronize moral expectations even when agents begin far apart.

S4.6. Ablation Studies. To isolate the contributions of trust-gating, belief updating, and virtual bargaining, we evaluate TGMI variants: (i) no-trust, (ii) no-belief, and (iii) no-bargaining. Each ablation reveals a specific failure mode.

Fig. S6. Fig. S4.6. Ablation analyses. (A) Removing trust leads to over-cooperation and vulnerability to exploitation. (B) Removing belief updating causes brittle misalignment with moral partners. (C) Removing bargaining collapses coordination in asymmetric games. This confirms that trust gating, CK-ToM inference, and virtual bargaining jointly constitute the mechanism enabling stable cooperation under moral uncertainty.

SI Dataset S1 (dataset_two.txt)

Type or paste legend here. Adding longer text to show what happens, to decide on alignment and/or indentations for multi-line or paragraph captions.

References