A Computational Theory of Trust-Gated Cooperation under Moral Uncertainty

Avi Sharma, Dasari Sai Harsh, Jainendra Shukla

IIIT-Delhi

This manuscript was compiled on December 13, 2025

Cooperation often hinges not only on predicting others' behavior but on understanding the moral standards they apply when evaluating outcomes. Yet most formal accounts of cooperation implicitly assume that partners share those standards, treating disagreements about fairness as noise. In real human groups—and increasingly in human–AI interactions—agents differ in how they prioritize efficiency, equality, or need, creating *moral uncertainty* that destabilizes reciprocity and leads to systematic misinterpretations of intent. We introduce *Trust-Gated Moral Inference* (TGMI), a computational model in which agents infer one another's moral priors, calibrate trust in whether these principles can support a jointly justifiable decision, and engage in virtual bargaining when sufficient overlap is inferred. Trust dynamically gates the influence of others' inferred norms on joint action selection, enabling cooperation when moral viewpoints are compatible and inducing a principled *moral fallback* to one's own fairness prior when they are not. We evaluate TGMI in a Moral Game Generator that varies both payoff structures and normative environments, as well as in classic social-dilemma settings. Analytically, we show that TGMI admits cooperative equilibria with bounded exploitability. Evolutionary simulations demonstrate that TGMI outperforms payoff-based and fairness-blind baselines across heterogeneous moral types and noise regimes. Together, these results identify trust-mediated moral inference as a general mechanism for sustaining cooperation across moral divides, offering a unified account of human cooperative reasoning and a foundation for socially aligned artificial agents.

moral uncertainty | trust-based inference | cooperation | game theory | cognitive science

Significance Statement

Human and artificial agents increasingly interact in environments where they hold different moral principles and cannot assume shared standards of fairness or justification. Such moral heterogeneity systematically affects how intentions are interpreted, when trust is extended, and whether cooperation is sustained. Standard reciprocity models—even those with theory of mind—struggle in these settings because they treat disagreement as noise rather than as substantive differences in moral commitments. We introduce a computational framework in which agents infer one another's moral priors, calibrate trust in whether these principles can ground a jointly justifiable decision, and engage in virtual bargaining when sufficient overlap is inferred. This trust-gated moral inference mechanism explains when cooperation succeeds across moral divides and when it predictably fails. The model provides both a theoretical account of human cooperative reasoning and a foundation for building socially aligned AI systems that can negotiate and cooperate across diverse normative perspectives.

Cooperation enables individuals—human or artificial—to pursue private goals while generating collective benefits (1–4). Classical accounts explain cooperation through mechanisms such as reciprocity, reputation, and repeated interaction (5, 6). These models have been remarkably successful, yet they rely on a tacit assumption: that partners evaluate outcomes using a shared moral or fairness metric (7, 8). When individuals disagree about what counts as fair, generous, or just, the same behavior can be interpreted in incompatible ways, causing well-intended cooperation to be mistaken for selfishness or exploitation (9–12).

Real societies display substantial moral pluralism (13, 14). People differ in the principles they prioritize—efficiency, equality, need, or rights—and partners may evaluate identical outcomes using divergent normative lenses (15–17). Bicchieri's account of social norms emphasizes that cooperation depends not only on behavioral expectations but also on shared *justificatory* expectations: agreement about the reasons that legitimate a joint action (18). When justificatory structures diverge, trust erodes, and behavior that is locally cooperative may nonetheless undermine social alignment.

Artificial agents face analogous challenges. In domains such as medical triage, content moderation, and multi-agent coordination, AI systems interact with partners whose moral commitments they neither share nor fully know (19–22). An AI that cooperates only with partners who share its value function is brittle; one that cooperates indiscriminately is easily exploited. Reasoning under *moral uncertainty*—uncertainty about others' moral priors—is therefore essential for socially aligned artificial agents.

Author affiliations: [1] IIIT-Delhi

Please provide details of author contributions here.

Please declare any competing interests here.

[1] A.O.(Author One) contributed equally to this work with A.T. (Author Two) (remove if not applicable).

[2] To whom correspondence should be addressed. E-mail: author.twoemail.com

Humans appear to navigate these challenges by forming structured beliefs about others' latent intentions and normative commitments. Computational models of social cognition formalize this ability as Bayesian theory of mind, in which agents infer partners' hidden utilities from sparse and noisy observations (23, 24). Yet existing approaches typically assume that agents differ only in cooperativeness, not in the underlying moral principles that guide their evaluations. When partners pursue substantively different notions of fairness or moral worth, such models mispredict both cooperation and its breakdown.

These considerations motivate a central question:

> **How can cooperation persist when agents are uncertain not only about others' intentions, but also about the moral principles that underlie those intentions?**

Addressing this question requires integrating three capacities that jointly structure human moral cognition: (i) inference over others' moral priors (25, 26); (ii) dynamically calibrated trust (27, 28); and (iii) mechanisms for forming jointly acceptable decisions under disagreement (29–32).

Here we develop a computational theory of cooperation under *moral uncertainty* that formalizes these ingredients. Our approach, **Trust-Gated Moral Inference (TGMI)**, rests on three components:

1. **Beliefs over others' moral priors**—probabilistic expectations about how partners weight fairness principles such as equality, efficiency, and need.

2. **A trust-modulated utility**, which adjusts how strongly an agent treats a partner's inferred moral values as admissible grounds for joint choice.

3. **Counterfactual moral negotiation**, implemented as a virtual bargaining process (33), which identifies actions mutually endorsable under the subset of moral principles that agents trust each other to share.

When trust is high, agents coordinate on a fairness-weighted joint utility steered by inferred shared principles; when trust is low, they revert to their intrinsic moral priors. In this view, cooperation becomes conditional on *normative compatibility* rather than behavioral mimicry alone (34, 35), capturing both the emergence of cooperation across moral divides and its predictable breakdown when shared justification cannot be established.

We first present the computational framework for cooperation under *moral uncertainty*, detailing how agents represent and update probabilistic beliefs about others' moral priors and how trust dynamics regulate the balance between moral and strategic reasoning (Fig. 1). We then derive the analytical properties of the model, establishing conditions for the existence, stability, and boundedness of cooperative equilibria (Fig. 2). Next, we introduce a *Moral Game Generator* extending prior game-generator frameworks (23) to construct a continuous-game environment that samples diverse social dilemmas in which agents vary not only in payoffs but also in moral-preference priors. This extension enables systematic evaluation of the model's robustness across environments differing in symmetry, noise, and value alignment (Fig. 3). Through agent-based simulations, we demonstrate that trust-gated moral inference sustains cooperation and collective welfare across heterogeneous partners and outperforms classical reciprocity and fairness-blind baselines (Fig. 4). Finally, we analyze emergent patterns of trust, belief convergence, and moral alignment, showing how counterfactual moral negotiation enables cooperation even when normative principles diverge (Fig. 5).

Together, these results provide a general framework for cooperation in morally heterogeneous societies and a foundation for designing artificial agents capable of aligning with partners whose values differ from their own.

Fig. 1. Belief and trust dynamics under moral uncertainty. **(A)** Each agent maintains probabilistic beliefs over a partner's moral priors—the relative weights they assign to fairness, efficiency, and need. Shown is one agent's prior distribution before interaction. **(B)** In a social-dilemma task, the donor allocates resources between self and other according to these inferred moral weights. **(C)** After observing the partner's choice, the agent updates both the posterior over the partner's moral priors and its trust variable τ. Actions consistent with shared principles (e.g., fair division) increase τ, whereas actions interpreted as self-interested reduce it. **(D)** The resulting learning loop: agents begin with prior beliefs B_0, act based on current trust and expected joint utility, observe others' behavior, compute its likelihood under competing moral models, update (B, τ), and repeat. This trust-gated inference process enables cooperation to adapt dynamically to moral diversity.

The Trust-Gated Moral Inference (TGMI) Model

We formalize cooperation under moral uncertainty as a repeated social decision-making process in which each agent must act while uncertain about both a partner's intentions and the moral principles guiding those intentions. A two-player game G consists of actions $(a_i, a_j) \in \mathcal{A}$, material payoffs $R_i(a_i, a_j)$, and a set of fairness principles $\mathcal{F} = \{\phi_1, \ldots, \phi_m\}$ with associated fairness mappings $F_\phi : \mathcal{A} \to [0, 1]$. Each agent i holds an *intrinsic moral prior* $B_i(\phi)$ over principles and a *moral-belief model* $\hat{B}_{i\to j}(\phi)$ representing its current estimate of partner j's fairness weights.

Moral utility. Agents evaluate outcomes through a combination of material reward and fairness assessments aggregated across principles. The relative weighting of intrinsic and inferred norms is governed by an *effective cooperation weight*

$$\kappa_i := \tau_i c_i,$$

where $\tau_i \in [0, 1]$ is a trust variable and $c_i = 1 - H(\hat{B}_{i\to j})/\log|\mathcal{F}|$ is an epistemic confidence term. The

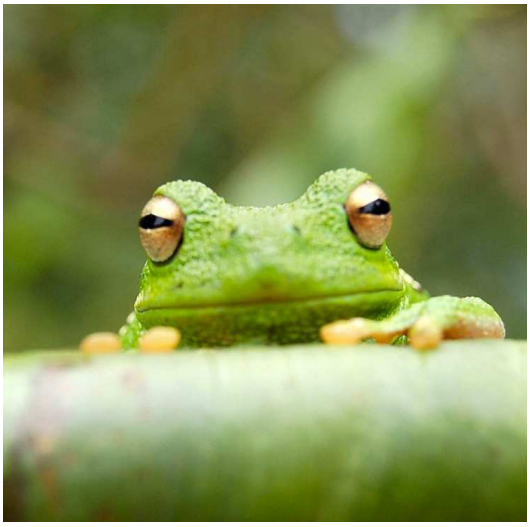

Fig. 2. The Moral Game Generator (MGG). Rather than repeating a single social dilemma, the Moral Game Generator samples an open-ended space of moral environments in which both the payoff structure and the underlying fairness norms vary probabilistically. The MGG is parameterized by a set of agents (colored circles), distributions over payoffs $R_i(a_i, a_j)$, and moral priors $\theta = (w_{\text{Max-Sum}}, w_{\text{Equal-Split}}, w_{\text{Rawls}})$ embedded within a continuous-action template. Each agent's moral prior θ_i is initialized from a symmetric Dirichlet distribution over fairness weights, providing a neutral yet diverse starting point across efficiency, equality, and need-based reasoning. Each sample $(G_1, \ldots, G_n)$ represents a unique moral decision-making problem combining a payoff landscape and a fairness environment drawn from $\mathcal{F} = \{\text{Max-Sum}, \text{Equal-Split}, \text{Rawls}\}$. In the illustration, the rotated agent on the left (e.g., player i in G_1) is the decision maker, selecting an action that determines both self-reward and fairness alignment with others. Together, these sampled interactions define a continuous distribution of moral dilemmas spanning efficiency–equity trade-offs, enabling TGMI to evaluate cooperation under moral heterogeneity.

resulting moral utility is

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

When $\kappa_i \to 0$, decisions reflect intrinsic fairness values; when $\kappa_i \to 1$, they align with the inferred fairness profile of the partner.

Counterfactual moral agreement via virtual bargaining. At each interaction, agents compute a counterfactual joint action representing what fair partners would endorse under mutual understanding. We adopt a virtual bargaining formulation, selecting a *Virtual Bargaining Equilibrium (VBE)*:

$$(a_i^{\text{VB}}, a_j^{\text{VB}}) = \arg\max_{a_i, a_j} \Big[(U_i(a_i, a_j) - d_i)_+ \Big]^{\gamma} \Big[(U_j(a_j, a_i) - d_j)_+ \Big]^{1-\gamma} \quad [2]$$

where $(x)_+ = \max(x, 0)$, $\gamma \in (0, 1)$ encodes bargaining asymmetry, and d_i, d_j are reservation utilities (initialized to 0). The VBE captures a counterfactual joint commitment rather than strategic best responses.

Fairness deviation and trust update. Agent i evaluates how far the partner's action deviates from what would have maximized i's intrinsic fairness evaluation:

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

$$d_i^{(t)} = \max_{a_j' \in \mathcal{A}_j} U_i^F(a_i^{\text{VB}}, a_j') - U_i^F(a_i^{\text{VB}}, a_j^{\text{VB}}). \quad [3]$$

Compliance is

$$s_i^{(t)} = \exp(-\lambda_{\text{dev}} d_i^{(t)}),$$

yielding a leaky-integration trust update:

$$\tau_i^{(t+1)} = (1-\eta)\tau_i^{(t)} + \eta \, s_i^{(t)}. \quad [4]$$

Common-Knowledge Theory of Mind (CK-ToM) belief update. To avoid infinite belief recursion, agents assume partner j updates from the same publicly observed interaction. Moral beliefs are updated by combining fairness evidence with intrinsic prior:

$$\tilde{B}_{i\to j}^{(t+1)}(\phi) = \hat{B}_{i\to j}^{(t)}(\phi) \, \exp\big(\beta \, \alpha \tau_i^{(t)} F_\phi(a_i^{\text{VB}}, a_j^{\text{VB}})\big) \, [B_i(\phi)]^{1-\alpha\tau_i^{(t)}}, \quad [5]$$

$$\hat{B}_{i\to j}^{(t+1)}(\phi) = \frac{\tilde{B}_{i\to j}^{(t+1)}(\phi)}{\sum_{\psi \in \mathcal{F}} \tilde{B}_{i\to j}^{(t+1)}(\psi)}. \quad [6]$$

Confidence updates via entropy reduction:

$$c_i^{(t+1)} = 1 - \frac{H(\hat{B}_{i\to j}^{(t+1)})}{\log |\mathcal{F}|}, \quad [7]$$

and the effective cooperation weight becomes

$$\kappa_i^{(t+1)} = \tau_i^{(t+1)} c_i^{(t+1)}.$$

Reservation utilities update to realized fairness:

$$d_i^{(t+1)} = U_i^F(a_i^{\text{VB}}, a_j^{\text{VB}}). \quad [8]$$

Emergence of cooperation. Eqs. 1–8 define a closed trust–belief–action loop. Agents beginning with divergent moral priors can converge to high trust and cooperative alignment. When partners act according to incompatible norms, trust collapses and behavior reverts to intrinsic fairness, yielding robustness against exploitation. Theoretical analysis (Appendix S1) establishes: (i) existence of at least one VBE fixed point, (ii) bounded exploitability, and (iii) stability of cooperative equilibria on a set of positive measure in the MGG. Full proofs appear in the Supplementary Appendix.

The Moral Game Generator (MGG)

Human and artificial agents rarely encounter the same social dilemma twice. The payoff structure, the incentives of partners, and the normative expectations that guide their judgments all vary from one interaction to the next. To evaluate TGMI across this diversity, we extend the Game Generator of Kleiman-Weiner (23) into a *Moral Game Generator (MGG)* that jointly samples both the material and moral structure of interactions.

Each sampled interaction specifies two components:

1. **Payoff landscape.** A material payoff function $R_i(a_i, a_j)$ is drawn from one of four continuous-action archetypes—*Dilemma*, *Assurance*, *Bargain*, and *Public-Goods*— capturing variation in incentives, externalities, and asymmetries.

2. **Moral environment.** Each agent receives a moral prior

$$B_i = (w_{\text{Max-Sum}}, \, w_{\text{Equal-Split}}, \, w_{\text{Rawls}})$$

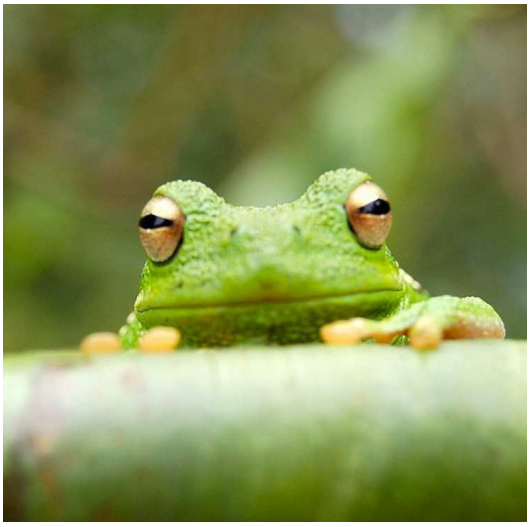

Fig. 3. Trust-Gated Moral Inference enables robust cooperation under moral diversity. (A) *Intragenerational belief and trust updating.* A probe TGMI agent interacted with three partner types—another TGMI agent (Left), a Fairness-Aligned agent (Center), and a Self-Interested agent (Right)—over 20 rounds. After each interaction, the agent updated its belief over the partner's moral priors and its trust variable τ. Dark curves show the mean of 1,000 runs; faint traces show individual trajectories. Across conditions, beliefs converge toward accurate moral inference, while trust selectively stabilizes with fair partners and decays with exploitative ones. (B–E) *Intergenerational dynamics in the Moral Game Generator (MGG) under repeated play with evolving populations.* (B) Steady-state abundance of TGMI agents as a function of game length T. (C) Steady-state abundance as a function of action error ε_a. (D) Combined trust–belief stability map: regions above the red line indicate majority (> 0.5) TGMI dominance in the steady-state population. (E) Mean population welfare across ε_a and T; higher TGMI abundance correlates with elevated joint payoffs due to resilient moral cooperation. Together, these results show that TGMI's trust-gated inference mechanism generalizes reciprocity to moral domains, producing cooperation robust to both behavioral noise and normative disagreement.

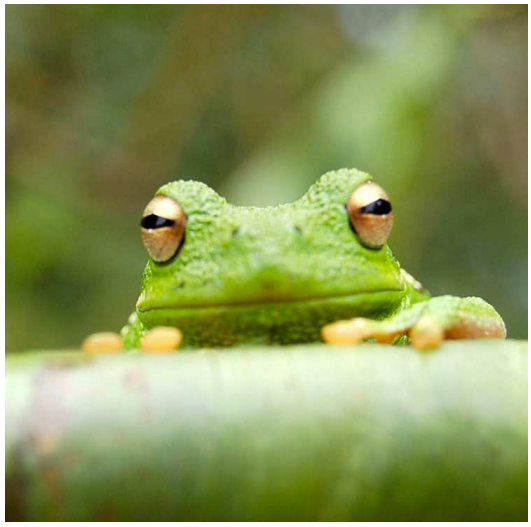

Fig. 4. Trust-Gated Moral Inference generalizes cooperation from limited social evidence. (A) *Intragenerational belief and trust updating* in a population of 10 TGMI agents, where each observes only a subset of others' interactions. Beliefs are shown after each observation for three partner types: another TGMI agent (Left), a Fairness-Aligned agent (Center), and a Self-Interested agent (Right). Dark curves show the average of 1,000 runs; faint lines denote individual trajectories. Across cases, agents learn to infer partners' moral priors and calibrate trust accordingly, sustaining cooperation even with sparse information. (B–D) *Intergenerational steady-state dynamics* of the population in the Moral Game Generator (MGG) under the Moran process, while varying (B) the probability of social observability ω, (C) the probability of action noise ε_a, and (D) the probability of perception noise ε_p. Bars represent the steady-state abundance of each agent type, averaged over 200 generations. TGMI maintains dominance across wide ranges of observability and action error but shows gradual sensitivity to misperceived fairness signals. (E–G) *Population welfare outcomes* for the scenarios in (B–D), showing that higher steady-state abundance of TGMI corresponds to greater collective payoff and fairness alignment. These results demonstrate that TGMI's trust-gated inference mechanism enables cooperation to persist even when agents must reason from limited or noisy moral observations.

sampled from a symmetric Dirichlet distribution. Fairness mappings translate payoffs into normative evaluations:

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

These mappings span utilitarian, egalitarian, and prioritarian principles, enabling agents with different priors to interpret the same material situation through distinct normative lenses.

All fairness values and payoffs are normalized to $[0, 1]$, ensuring comparability across games and allowing TGMI to operate on a smooth joint material–moral landscape.

The MGG therefore samples a high-dimensional space of social dilemmas differing in both incentives and moral structure. Whereas the original generator varied only payoffs, the MGG introduces systematic variation in the *value systems* that agents expect others to follow. This is essential for evaluating TGMI: moral uncertainty arises only when fairness environments and moral priors fluctuate in ways that agents must infer and adapt to.

Finally, interaction-level noise is incorporated through independent action errors (ε_a) and optional perception noise on fairness evaluations (ε_p). Full parameterizations, sampling procedures, and robustness checks appear in the *Supplementary Appendix*.

Materials and Methods

Trust-Gated Moral Inference (TGMI). TGMI integrates three computational components: (i) trust–confidence–gated moral utility (Eq. 2), (ii) counterfactual joint action selection via Virtual Bargaining (Eq. 3), and (iii) partner–model inference using a Common-Knowledge Theory of Mind update (Eq. 4). At each round, agents evaluate actions through their intrinsic moral prior B_i, their beliefs about a partner's norms $\hat{B}_{i\to j}$, and a dynamic cooperation weight $\kappa_i = \tau_i c_i$ that combines trust and epistemic confidence. Trust increases when a partner behaves in ways consistent with i's moral expectations and decreases otherwise; beliefs adjust through a likelihood-weighted mixture of observed fairness and self-anchoring.

Algorithm 1 summarizes the computational loop. Mathematical definitions of utility, bargaining, trust updating, and CK-ToM inference appear in Eqs. 1–4 and are elaborated in Appendix S1.

Moral Game Generator (MGG). To evaluate cooperation under joint variability in incentives and moral values, we extend the Game Generator of Kleiman-Weiner *et al.* (2025) into a **Moral Game Generator (MGG)**. Each sampled game G specifies both a material payoff landscape and a moral environment.

Payoff structure. For each interaction, a continuous-action payoff function $R_i(a_i, a_j) \in [0, 1]$ is drawn from one of four archetypes commonly used to model social dilemmas: *Dilemma*, *Assurance*,

Algorithm 1 Trust-Gated Moral Inference (TGMI) learning loop for agent i

Initialize intrinsic moral prior B_i, partner-belief $\hat{B}_{i\to j} \sim \text{Dirichlet}(1,\dots,1)$, trust τ_i, confidence c_i, cooperation weight $\kappa_i = \tau_i c_i$, and reservation value $d_i = 0$.

for $t = 1$ to T **do**

(a) Moral utility evaluation

  for each (a_i, a_j) in the action grid **do**

    Compute agent i's trust-gated moral utility $U_i(a_i, a_j)$ (Eq. 2).

    Compute agent i's internal model of partner j's utility $\widehat{U}_j^{(i)}(a_j, a_i)$ using $\hat{B}_{i\to j}$ (Eq. 2 with partner weights).

(b) Virtual Bargaining (counterfactual coordination)

  Select joint action $(a_i^{\text{VB}}, a_j^{\text{VB}})$ by solving the Nash-style product (Eq. 3):

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

(c) Fairness deviation and trust update

  Compute fairness-only utility $U_i^F(a_i, a_j)$ (Eq. Fairness).

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

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

(d) CK-ToM belief update and confidence

  For each fairness norm $\phi \in \mathcal{F}$, form the unnormalized update:

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

  Normalize: $\hat{B}_{i\to j}^{(t+1)}(\phi) = \tilde{B}_{i\to j}^{(t+1)}(\phi) / \sum_{\phi'} \tilde{B}_{i\to j}^{(t+1)}(\phi')$.

  Update confidence $c_i^{(t+1)} = 1 - H(\hat{B}_{i\to j}^{(t+1)}) / \log|\mathcal{F}|$.

  Update cooperation weight $\kappa_i^{(t+1)} = \tau_i^{(t+1)} c_i^{(t+1)}$.

  Update reservation value $d_i^{(t+1)} = U_i^F(a_i^{\text{VB}}, a_j^{\text{VB}})$.

Bargain, and *Public-Goods*. These surfaces define the underlying incentive structure and allow for asymmetric or noisy payoffs.

Moral environment. Each agent is assigned a moral prior

$$B_i = (w_{\text{Max-Sum}}, w_{\text{Equal-Split}}, w_{\text{Rawls}})$$

sampled from a Dirichlet distribution over fairness principles

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

Fairness mappings translate material payoffs into normative evaluations:

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

All values are normalized to $[0, 1]$, ensuring comparability across material and fairness-based utilities.

Moral heterogeneity sweep. To systematically vary the diversity of moral preferences in the population, we parameterize the Dirichlet prior with a concentration parameter α_{dir}:

$$B_i \sim \text{Dirichlet}(\alpha_{\text{dir}} \cdot \mathbf{1}).$$

Low α_{dir} yields sharply peaked, idiosyncratic moral priors; high α_{dir} produces convergence around the uniform prior. This sweep enables direct tests of when TGMI succeeds or fails as moral diversity increases, probing cooperation under moral uncertainty.

Noise. Each round introduces independent *action errors* (ϵ_a) and optional *perception noise* (ϵ_p) applied to fairness evaluations F_ϕ. After every interaction, a new payoff–fairness pair is sampled, generating an open-ended sequence of heterogeneous dilemmas.

Full implementation details, parameter sweeps, and sampling code are provided in the *Supplementary Appendix*.

Baseline and Comparison Agents. To assess the distinct contribution of trust-gated moral inference, we compare TGMI with a broad set of reciprocal and automaton strategies widely studied in evolutionary game theory. These baselines span unconditional cooperation, strict reciprocity, generous reciprocity, error correction, and exploitative behavior, and include **Tit-for-Tat (TFT)**, **Generous TFT**, **Win–Stay–Lose–Shift (WSLS)**, **Forgiver**, **Always Cooperate (AllC)**, and **Always Defect (AllD)**. Behavioral rules follow canonical definitions in (1, 3). Together, these strategies constitute a comprehensive set of non-moral baselines relying solely on behavioral reciprocity.

We also evaluate three targeted **TGMI ablations** designed to isolate the components required for cooperation under moral uncertainty: (1) *no-trust*, in which the trust variable is held fixed and interactions are treated as norm-neutral; (2) *no-belief*, in which agents do not update their estimates of partners' moral priors; and (3) *no-bargaining*, in which virtual bargaining is removed and agents optimize only their own moral utility. These ablations clarify how trust gating, moral inference, and counterfactual coordination jointly sustain cooperation.

Evolutionary Analysis in the MGG. We assess the long-run stability of cooperation using a finite-population Moran process. A population of $N = 10$ agents interacts for T rounds in the MGG; after each episode, social learning updates strategy frequencies based on relative performance.

Agent types. We compare TGMI with a set of classical reciprocal and automaton strategies:

$$\mathcal{T} = \{\text{TGMI, TFT, GTFT, WSLS, Forgiver, AllC, AllD}.$$

TGMI ablations remove specific components—trust gating, belief updating, or virtual bargaining—to isolate their functional contributions.

Returns and selection. For an agent i of type k, the fairness-weighted return is

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

where $\omega \in [0, 1]$ interpolates between purely material and morally weighted selection. Expected returns Φ_k across $S = 200$ replicates are mapped to copying probabilities via a softmax rule with selection strength s. Mutation at rate $\mu = 10^{-3}$ enables exploration of the strategy space. The stationary distribution is obtained as the normalized left eigenvector of the transition matrix.

Moral heterogeneity and cooperation. Evolutionary simulations are repeated over a grid of Dirichlet concentrations α_{dir}, action noise ϵ_a, perception noise ϵ_p, and observability q. This allows us to map how TGMI's evolutionary abundance and population-level welfare change as moral diversity increases. The resulting "heterogeneity–cooperation landscape" shows that TGMI maintains positive abundance and high welfare across a broad region of parameter space, even when agents' moral priors differ sharply.

All parameters, convergence diagnostics, and full code are provided in the *Supplementary Appendix*.

ACKNOWLEDGMENTS. Please include your acknowledgments here, set in a single paragraph. Please do not include any acknowledgments in the Supporting Information, or anywhere else in the manuscript.

1. R Axelrod, WD Hamilton, The evolution of cooperation. *Science* **211**, 1390–1396 (1981).
2. MA Nowak, K Sigmund, A strategy of win–stay, lose–shift that outperforms tit-for-tat in the prisoner's dilemma game. *Nature* **364**, 56–58 (1993).
3. MA Nowak, Five rules for the evolution of cooperation. *Science* **314**, 1560–1563 (2006).
4. DG Rand, MA Nowak, Human cooperation. *Trends Cogn. Sci.* **17**, 413–425 (2013).
5. RL Trivers, The evolution of reciprocal altruism. *The Q. review biology* **46**, 35–57 (1971).
6. H Ohtsuki, Y Iwasa, The leading eight: social norms that can maintain cooperation by indirect reciprocity. *J. theoretical biology* **239**, 435–444 (2006).
7. M Rabin, Incorporating fairness into game theory and economics. *The Am. economic review* pp. 1281–1302 (1993).
8. E Fehr, KM Schmidt, A theory of fairness, competition, and cooperation. *The quarterly journal economics* **114**, 817–868 (1999).
9. J Rawls, *A Theory of Justice.* (Harvard University Press), (1971).
10. A Sen, The idea of justice. *J. human development* **9**, 331–342 (2008).
11. M Fleurbaey, Fairness, responsibility and welfare. *Book Manuscr.* (2007).
12. M MacAskill, K Bykvist, T Ord, *Moral uncertainty.* (Oxford University Press), (2020).
13. I Berlin, *The crooked timber of humanity: Chapters in the history of ideas.* (Princeton University Press), (2013).
14. DB Wong, *Natural moralities: A defense of pluralistic relativism.* (Oxford University Press), (2009).
15. MJ Sandel, Political liberalism (1994).
16. OS Curry, DA Mullins, H Whitehouse, Is it good to cooperate? testing the theory of morality-as-cooperation in 60 societies. *Curr. anthropology* **60**, 47–69 (2019).
17. F Cushman, Rationalization is rational. *Behav. Brain Sci.* **43**, e28 (2020).
18. C Bicchieri, *The grammar of society: The nature and dynamics of social norms.* (Cambridge University Press), (2005).
19. I Gabriel, Artificial intelligence, values, and alignment. *Minds machines* **30**, 411–437 (2020).
20. I Rahwan, et al., Machine behaviour. *Nature* **568**, 477–486 (2019).
21. A Dafoe, et al., Open problems in cooperative ai. *arXiv preprint arXiv:2012.08630* (2020).
22. T Swoboda, L Lauwaert, Can artificial intelligence embody moral values? *AI Ethics* pp. 1–12 (2025).
23. M Kleiman-Weiner, A Vientos, DG Rand, JB Tenenbaum, Evolving general cooperation with a bayesian theory of mind. *Proc. Natl. Acad. Sci.* **122**, e2400993122 (2025).
24. CL Baker, J Jara-Ettinger, R Saxe, JB Tenenbaum, Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nat. Hum. Behav.* **1**, 0064 (2017).
25. M Kleiman-Weiner, R Saxe, JB Tenenbaum, Learning a commonsense moral theory. *Cognition* **167**, 107–123 (2017).
26. J Jara-Ettinger, Theory of mind as inverse reinforcement learning. *Curr. Opin. Behav. Sci.* **29**, 105–110 (2019).
27. AM Evans, JI Krueger, The psychology (and economics) of trust. *Soc. Pers. Psychol. Compass* **3**, 1003–1017 (2009).
28. I Thielmann, BE Hilbig, Trust: An integrative review from a person–situation perspective. *Rev. Gen. Psychol.* **19**, 249–277 (2015).
29. C Hilbe, Š Šimsa, K Chatterjee, MA Nowak, Evolution of cooperation in stochastic games. *Nature* **559**, 246–249 (2018).
30. JJ Jordan, M Hoffman, P Bloom, DG Rand, Third-party punishment as a costly signal of trustworthiness. *Nature* **530**, 473–476 (2016).
31. M Tomasello, M Carpenter, J Call, T Behne, H Moll, Understanding and sharing intentions: The origins of cultural cognition. *Behav. Brain Sci.* **28**, 675–691 (2005).
32. S Hadfield-Menell, S Russell, P Abbeel, A Dragan, Cooperative inverse reinforcement learning. pp. 3909–3917 (2016).
33. JB Misyak, N Chater, Virtual bargaining: A theory of social decision-making. *Philos. Transactions Royal Soc. B* **369**, 20130487 (2014).
34. RX Hawkins, ND Goodman, RL Goldstone, The emergence of social norms and conventions. *Trends cognitive sciences* **23**, 158–169 (2019).
35. S Levine, M Kleiman-Weiner, L Schulz, J Tenenbaum, F Cushman, The logic of universalization guides moral judgment. *Proc. Natl. Acad. Sci.* **117**, 26158–26169 (2020).