A Computational Theory of Cooperation under Moral Uncertainty

Jainendra Shukla

IIIT-Delhi

This manuscript was compiled on November 15, 2025

Theories of cooperation explain how self-interested agents achieve mutual benefit through reciprocity and reputation. However, these approaches typically assume that all agents evaluate outcomes according to shared moral preferences—an assumption rarely met in real societies. When moral beliefs diverge or remain uncertain, standard reciprocity fails because agents may misinterpret fair behavior as defection. We develop the *Trust-Gated Moral Inference (TGMI)* model of cooperation under *moral uncertainty*, in which each agent maintains probabilistic beliefs about others' moral priors—such as fairness or equality—and regulates cooperation through a trust-based control signal that determines how strongly to weight moral versus strategic reasoning. When trust is high, agents optimize a shared fairness utility; when trust is low, they revert to self-interest. In analytical and simulated social-dilemma environments, this framework sustains cooperation across diverse moral types, limits exploitation, and converges to stable cooperative equilibria. Overall, TGMI formalizes the advantage of *trust-mediated moral reasoning* for sustaining cooperation under moral uncertainty and provides a foundation for socially aligned artificial agents capable of cooperating across moral diversity.

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

Cooperation is one of the most powerful achievements of intelligent systems. It enables individuals—human or artificial—to pursue private goals while generating collective benefits (? ? ? ?). Decades of research show how mechanisms such as reciprocity, reputation, and repeated interaction can stabilize cooperation even among self-interested agents (? ? ? ? ? ?). Yet these frameworks rely on a hidden assumption: that all participants evaluate outcomes according to a shared moral metric (? ? ? ?). When agents disagree about what counts as fair or acceptable, classical models of cooperation lose traction (? ? ? ? ?).

Real societies are morally pluralistic (? ? ? ?). People differ in the principles they prioritize—valuing, for instance, equality, efficiency, or the welfare of the least advantaged (? ?). Even prosocial intentions can be misinterpreted when moral standards diverge (? ? ? ?). Artificial agents face a similar dilemma. When deployed in social or decision-making settings—allocating medical resources, moderating online interactions, or coordinating autonomous systems—they must act among partners whose moral beliefs and goals may not align (? ? ? ? ?). In such environments, the challenge is not merely predicting others' behavior, but reasoning about the *normative models* that guide that behavior: what others believe is right, and how much those beliefs can be trusted (? ?).

People learn about cooperative partners and their motives by integrating observations into mental models of others' latent utilities—a process formalized in recent computational accounts of reciprocity grounded in theory of mind (?). While such models infer others' payoff functions, they typically assume a shared moral metric. We generalize this approach to settings in which the moral priors themselves are uncertain or divergent.

Significance Statement

Moral uncertainty arises when individuals are unsure which principles of fairness or justice others follow. Most models of cooperation assume shared moral values, leaving open the question of how cooperation can persist when those values diverge. We develop a computational theory of cooperation under *moral uncertainty*, in which agents infer others' moral preferences and regulate cooperation through *trust-based moral inference*. By learning when to act strategically and when to act morally, agents sustain cooperation even with partners who reason differently. In evolutionary social-dilemma simulations, this process maintains cooperation across diverse moral types while resisting exploitation. These results suggest that *trust-mediated moral reasoning* provides a general mechanism for aligning behavior across moral diversity.

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

This challenge motivates a central question: **How can cooperation persist under moral uncertainty**—when agents are uncertain not only about others' intentions, but about the moral principles those intentions express? Addressing this question requires moving beyond standard tools of reciprocity or reward maximization (? ? ? ?). Agents must infer the moral values shaping others' choices (? ? ?), adjust their willingness to cooperate as trust evolves (? ? ?), and remain robust to disagreement and noise (? ? ?). These capacities mirror human moral cognition (?), where trust and moral inference jointly shape social alignment. Capturing these dynamics computationally is essential for developing artificial systems that can coexist within diverse moral ecologies (? ?).

We propose a computational theory of cooperation under *moral uncertainty* that formalizes this process (? ? ?). Each agent maintains probabilistic beliefs about others' moral priors—how they weigh fairness, efficiency, or need—and updates these beliefs through interaction (? ? ?). A trust variable modulates whether decisions follow instrumental self-interest or a fairness-weighted joint utility (? ?). When trust is high, agents optimize outcomes consistent with inferred shared principles; when trust is low, they revert to self-interest. Cooperation thus becomes conditional on *moral consistency* rather than mere behavioral reciprocity: agents cooperate with those whose actions can be interpreted as following mutually endorsable norms (? ? ?).

Analytically, we show that this framework yields cooperative equilibria that are stable and bounded against exploitation, even when agents' moral priors differ. Using a continuous-game generator spanning symmetric, asymmetric, and noisy environments, we demonstrate that *trust-gated moral inference* sustains cooperation and collective welfare beyond what is achieved by fairness-blind, purely reciprocal, or fixed-rule baselines. Conceptually, the mechanism can be interpreted as a form of counterfactual moral negotiation—akin to a *virtual bargaining equilibrium* (?)—in which agents implicitly reason about the agreements they would reach under mutual understanding. Together, these results offer a foundation for modeling cooperation in heterogeneous moral environments and for building socially aligned artificial agents that deliberate not only over what others do, but over what they ought reasonably to endorse (?).

We develop the Trust-Gated Moral Inference (TGMI) model, which formalizes cooperation under moral uncertainty as a process of belief-driven moral inference modulated by evolving trust. 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 adapt the Game Generator framework introduced by Kleiman-Weiner *et al.* (2025) 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).

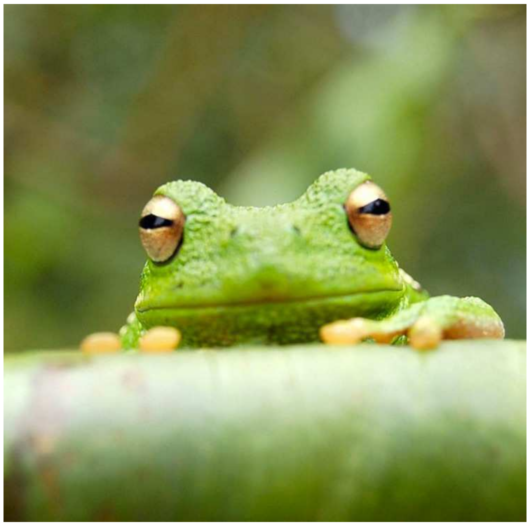

Fig. 1. Belief and trust dynamics under moral uncertainty. (A) Each agent maintains probabilistic beliefs over a partner's moral priors—the relative weights they assign to fairness, efficiency, and need. Shown is one agent's prior distribution before interaction. **(B)** In a social-dilemma task, the 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

A game G consists of players $i \in N$, joint actions $a = (a_i, a_j) \in \mathcal{A}$, and payoffs $R_i(a_i, a_j)$. Let $\mathcal{A} = \mathcal{A}_i \times \mathcal{A}_j$ and a_{-i} denote the actions of all other players. Player i receives payoff $R_i(a_i, a_j)$ and evaluates it through a private *moral utility function* $U_i(a_i, a_j \mid B_i, \tau_i)$, where B_i denotes its *moral prior* over fairness principles and $\tau_i \in [0, 1]$ the *trust* variable. Agents choose the action that maximizes expected utility:

$$a_i^* = \arg\max_{a_i \in \mathcal{A}_i} U_i(a_i, a_{-i}). \qquad [1]$$

Players occasionally commit action errors with probability ϵ_a; only realized actions are observed.

Notation. For brevity, a joint action profile is denoted by $a = (a_i, a_j)$, where $a_i \in \mathcal{A}_i$ and $a_j \in \mathcal{A}_j$. Each fairness mapping $F_\phi : \mathcal{A}_i \times \mathcal{A}_j \to [0, 1]$ evaluates the fairness of joint outcomes under principle $\phi \in \mathcal{F}$. When the arguments are implicit, we write $F_\phi(a)$ as shorthand for $F_\phi(a_i, a_j)$. In time-indexed updates, such as the CK-ToM belief update or simulation loop, $F_\phi(a_i^{(t)}, a_j^{(t)})$ denotes evaluation at the realized actions on round t.

The key innovation of TGMI is that utility depends on a single *effective cooperation weight* $\kappa_i := \tau_i c_i \in [0, 1]$, which jointly

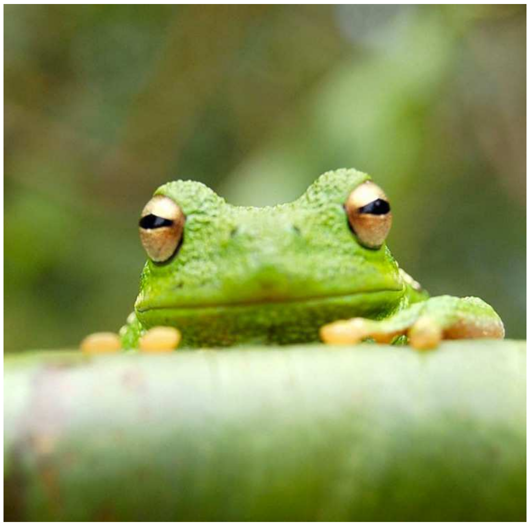

Fig. 2. The Moral Game Generator (MGG). Rather than repeating a single social dilemma, the Moral Game Generator samples an open-ended space of moral environments in which both the payoff structure and the underlying fairness norms vary probabilistically. The MGG is parameterized by a set of agents (colored 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 $\mathcal{F} = \{\text{Max-Sum, Equal-Split, Rawls}\}$. 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.

captures how much player i *trusts* a partner's intentions (τ_i) and how *confidently* it understands their moral model (c_i). Player i values not only the fairness of outcomes expected from morally aligned partners but also remains anchored in its own intrinsic moral standards when alignment is uncertain. Formally, the trust–confidence–gated moral utility is defined as:

$$U_i(a) = \sum_{\phi \in \mathcal{F}} \Big[(1-\kappa_i)\, B_i(\phi) + \kappa_i\, \hat{B}_{i\to j}(\phi)\Big] F_\phi(a) \qquad [2]$$

where $B_i(\phi)$ denotes player i's moral prior over fairness principles $\mathcal{F} = \{\text{Max-Sum, Equal-Split, Rawls, Prioritarian}\}$, and $\hat{B}_{i\to j}(\phi)$ its belief about partner j's moral norms inferred through interaction. Each fairness function $F_\phi : \mathcal{A}_i \times \mathcal{A}_j \to [0,1]$ maps joint outcomes to fairness under principle ϕ, quantifying efficiency, equality, or need.

The parameter κ_i acts as a dynamic moral gate: when $\kappa_i \to 0$, decisions reflect self-evaluated moral fairness (purely intrinsic reasoning); when $\kappa_i \to 1$, decisions align fully with the inferred partner norms, yielding joint moral reasoning. Trust evolves according to

$$\tau_i^{(t+1)} = (1-\eta)\tau_i^{(t)} + \eta\, s_i^{(t)},$$

where $s_i^{(t)}$ denotes observed compliance or fairness consistency, while confidence is updated by

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

with $H(\cdot)$ the Shannon entropy of the belief distribution. Together, these updates produce smooth variations in κ_i over time, allowing TGMI agents to modulate cooperation continuously as trust and epistemic certainty coevolve. A detailed derivation of the joint update and its equilibrium properties is provided in the *Supplementary Appendix*.

Unlike reciprocity-based models that revert to self-interest when trust collapses, TGMI maintains a *moral fallback*: when $\tau_i \to 0$ or $c_i \to 0$, the agent acts according to its intrinsic moral prior $B_i(\phi)$ rather than the material payoff $R_i(a)$. This ensures that even under distrust, behavior remains consistent with the agent's moral identity while ceasing to depend on uncertain partner alignment.

In social and artificial environments, cooperative decisions often require reconciling heterogeneous values rather than pursuing identical goals. TGMI focuses on the portion of agents' utilities that governs how they value others' welfare—abstracting away from idiosyncratic or non-social preferences. By reasoning about fairness trade-offs across moral principles, agents approximate the agreement that mutually fair partners would jointly endorse. At each round, agents engage in *virtual bargaining*—a process of counterfactual reasoning that approximates the agreement fair partners would reach under full mutual trust (? ?). They select a joint action that maximizes a Nash-style product of moral gains above reservation utilities:

$$(a_i^{\text{VB}}, a_j^{\text{VB}}) = \arg\max_{a_i, a_j} \Big[(U_i(a_i,a_j) - d_i)_+\Big]^{\gamma} \Big[(U_j(a_j,a_i) - d_j)_+\Big]^{1-\gamma} \qquad [3]$$

where $(x)_+ = \max(x, 0)$, bargaining asymmetry $\gamma \in (0,1)$ (typically $\gamma = \frac{1}{2}$), and d_i, d_j are reservation utilities taken from the previous round's fairness values (neutral defaults at $t = 1$). The resulting *Virtual Bargaining Equilibrium* (VBE) is approximately incentive-compatible (ϵ–Nash consistent) and captures the counterfactual agreement both partners would endorse if trust were complete.

When $\tau_i \to 1$, TGMI reduces to joint moral optimization; when $\tau_i \to 0$, it converges to a self-interested best response—paralleling the *Selfish* and *Altruistic* baselines in classical cooperation models. Virtual bargaining thus generalizes reciprocity: rather than mirroring observed actions, agents negotiate over the principles that justify them.

To model how agents infer one another's moral reasoning, TGMI builds on the Bayesian theory-of-mind (BToM) framework (? ? ?), but adapts it into a *Common-Knowledge Theory of Mind* (CK-ToM) formulation (? ?) that avoids recursive belief nesting. This formulation conditions on the publicly shared interaction history, enabling efficient higher-order reasoning without exponential cognitive cost.

Trust evolves through experience: after observing a partner's action, player i measures *fairness deviation*—the gap between observed and ideal moral outcomes—and updates trust asymmetrically. Beliefs about partner norms are then updated using the CK-ToM rule:

$$\hat{B}_{i\to j}^{(t+1)}(\phi) \propto \hat{B}_{i\to j}^{(t)}(\phi) \Big[\exp\big(\beta F_\phi(a_i^{(t)}, a_j^{(t)})\big)\Big]^{1-\alpha\tau_i^{(t)}} \Big[B_i(\phi)\Big]^{\alpha\tau_i^{(t)}}, \qquad [4]$$

where $\beta > 0$ controls sensitivity to fairness evidence, and $\alpha \in [0,1]$ regulates self-anchoring under trust. The proportional update is normalized so that $\sum_\phi \hat{B}^{(t+1)}_{i\to j}(\phi) = 1$. High trust anchors beliefs in the partner's observed fairness, whereas low trust reverts them toward the self prior. This CK-ToM update thus formalizes mutual moral inference without the infinite regress of nested belief models.

Together, Eqs. 1–4 define a closed learning loop: agents infer others' moral norms, act cooperatively through virtual bargaining, observe deviations, update trust, and revise beliefs. At initialization, each agent holds a fixed moral prior $B_i(\phi)$ representing its intrinsic fairness weights, while its belief about a partner's norms, $\hat{B}_{i\to j}(\phi)$, is drawn from a symmetric Dirichlet distribution over the fairness space $\mathcal{F}$, ensuring neutral expected weights ($\mathbb{E}[\hat{B}_{i\to j}] = 1/n$). This provides a neutral yet probabilistic starting point—agents begin uncertain about others' moral orientations but without bias toward selfishness or fairness—and results remain robust across different hyperparameter choices.

Through this mechanism, cooperative behavior emerges naturally: trust increases with consistent fairness, collapses under exploitation, and recovers when moral alignment is re-established. Overall, TGMI yields stable cooperative equilibria that are bounded against exploitation and recover gracefully after temporary defections—advancing a *computational theory of cooperation under moral uncertainty*, in which agents deliberate not only over what others do, but over what all could reasonably endorse.

Full derivations, update rules for moral fallback, and simulation details are provided in the *Supplementary Appendix*. Theoretical analysis (Appendix S1) demonstrates that the TGMI learning dynamics admit at least one *Virtual Bargaining Equilibrium (VBE)*, in which no agent can unilaterally increase its trust-gated utility. Moreover, deviations in realized fairness and payoff are bounded above by a finite constant, ensuring the stability and robustness of cooperative equilibrium behavior.

The Moral Game Generator (MGG)

Building on the Game Generator introduced by Kleiman-Weiner *et al.* (2025), we develop a **Moral Game Generator (MGG)** that extends variation from purely material to moral domains. Whereas the original generator sampled only payoff configurations, the MGG jointly samples both **payoff structures** and **moral priors**, producing continuous-action dilemmas that vary simultaneously in incentives, fairness expectations, and normative alignment.

Each sampled interaction specifies:

1. a payoff surface $R_i(a_i, a_j)$ drawn from one of several archetypes—*Dilemma*, *Bargain*, *Assurance*, or *Public-Goods*—defining the material incentive landscape; and

2. a moral environment defined by fairness principles $\mathcal{F} = \{\text{Max-Sum}, \text{Equal-Split}, \text{Rawls}\}$, where each $F_\phi(a_i, a_j)$ quantifies fairness under utilitarian, egalitarian, or prioritarian reasoning, respectively.

Formally, the fairness mappings are defined as

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

Each function $F_\phi(a_i, a_j) \in [0, 1]$ evaluates the fairness of joint outcomes under principle ϕ, allowing agents with different moral priors $B_i(\phi)$ to perceive the same material situation through distinct normative lenses.

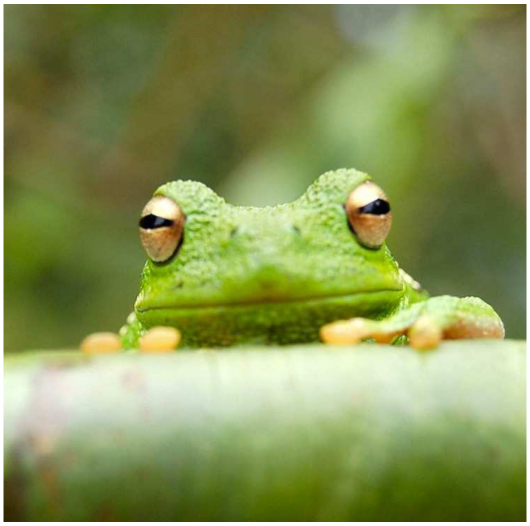

Fig. 3. Trust-Gated Moral Inference enables robust cooperation under moral diversity. (A) *Intragenerational belief and trust updating.* A probe TGMI agent interacted with three partner types—another TGMI agent (Left), a Fairness-Aligned agent (Center), and a Self-Interested agent (Right)—over 20 rounds. 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.

All payoffs and fairness utilities are normalized to $[0, 1]$, yielding a continuous landscape in which cooperation and fairness can diverge. By systematically varying both payoff asymmetry and moral divergence, the MGG enables TGMI to be evaluated for its **robustness to moral heterogeneity** rather than mere sensitivity to material uncertainty. Full parameterizations, sampling procedures, and additional fairness formulations are provided in the *Supplementary Appendix*.

Materials and Methods

Trust-Gated Moral Inference (TGMI) Model. Algorithm 1 specifies the operational loop of the Trust-Gated Moral Inference (TGMI) agent described by Eqs. 1–4. Each agent maintains probabilistic beliefs $\hat{B}_{i\to j}$ over a partner's moral priorities and a scalar trust variable τ_i that regulates the balance between joint cooperation and principled self-consistency. At every round, the agent computes its moral utility $U_i(a)$ using the effective cooperation weight $\kappa_i = \tau_i c_i$ (Eq. 2), selects a counterfactual joint action $(a_i^{(\text{VB})}, a_j^{(\text{VB})})$ through virtual bargaining (Eq. 3), and updates beliefs and trust according to the CK-ToM rule (Eq. 4).

Unlike reciprocity-based models that revert to self-interest when trust collapses, TGMI enforces a *moral fallback*: as $\tau_i \to 0$ or confidence $c_i \to 0$, the agent defaults to its intrinsic moral prior $B_i(\phi)$ rather than to material payoff maximization. This guarantees ethical coherence under uncertainty while allowing smooth transitions between mutual moral cooperation and principled independence.

The learning loop alternates through four stages—(a) moral decision-making via trust-gated utility, (b) observation and fairness deviation, (c) asymmetric trust updating, and (d) belief revision

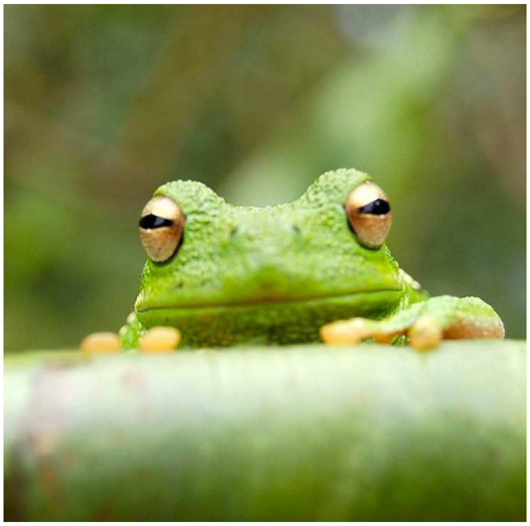

Fig. 4. Trust-Gated Moral Inference generalizes cooperation from limited social evidence. (A) *Intragenerational belief and trust updating* in a population of 10 TGMI agents, where each observes only a subset of others' interactions. Beliefs are shown after each observation for three partner types: another 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.

through CK-ToM inference—jointly defining how agents infer, align, and adapt moral expectations over repeated interactions. For reproducibility and welfare analysis, each iteration logs $\{\tau_i^{(t)}, c_i^{(t)}, \kappa_i^{(t)}, d_i^{(t)}, U_i(a_i^{(\mathrm{VB})}, a_j^{(\mathrm{VB})}), U_i^F(a_i^{(\mathrm{VB})}, a_j^{(\mathrm{VB})}), R_i(a_i^{(\mathrm{VB})}, a_j^{(\mathrm{VB})}), \hat{B}_{i\to j}^{(t)}\}$, enabling trajectory-level tracking of trust, belief convergence, and cooperative stability.

Together, these components define the TGMI learning loop that governs the emergence and stability of cooperation under moral uncertainty.

Moral Game Generator (MGG). The Moral Game Generator (MGG) implements the sampling process described in the main text (Eqs. 2–3), extending the Game Generator framework of Kleiman-Weiner *et al.* (2025) to incorporate moral variability. Each sample G_i defines a continuous-action game by specifying both a payoff surface and a moral environment. For each interaction, the number of players is set to two for intragenerational learning and two or three for evolutionary analyses. A payoff function $R_i(a_i, a_j)$ is drawn from one of four archetypes—**Dilemma**, **Assurance**, **Bargain**, or **Public-Goods**—each defining a continuous-action environment with normalized payoffs $R_i \in [0, 1]$.

Each agent i is assigned a moral prior $B_i = (w_{\text{Max-Sum}}, w_{\text{Equal-Split}}, w_{\text{Rawls}})$ sampled from a symmetric Dirichlet distribution, providing neutral diversity across fairness principles $\mathcal{F} = \{\text{Max-Sum}, \text{Equal-Split}, \text{Rawls}\}$. Fairness mappings follow those defined in the main text (Eq. 2): $F_{\text{Max-Sum}} = (R_i + R_j)/R_{\max}$, $F_{\text{Equal-Split}} = 1 - |R_i - R_j|/R_{\max}$, $F_{\text{Rawls}} = \min(R_i, R_j)/R_{\max}$.

These functions generate heterogeneous moral landscapes while maintaining consistent normalization across agents and trials.

Each game includes an independent **action error** ϵ_a (random substitution of the chosen action) and, optionally, a **perception**

Algorithm 1 Trust-Gated Moral Inference (TGMI) Learning and Update Loop for Agent i

```
Input: Fairness principles $\mathcal{F}$; payoff function $R_i(a_i, a_j)$; intrinsic moral prior $B_i(\phi)$
Initialize:
Belief over partner's norms $\hat{B}_{i\to j}(\phi) \sim \mathrm{Dirichlet}(1, \dots, 1)$
    ▷ Symmetric prior ⇒ $\mathbb{E}[\hat{B}] = 1/|\mathcal{F}|$
Trust $\tau_i \leftarrow \tau_0$, confidence $c_i \leftarrow 1 - H(\hat{B}_{i\to j})/\log|\mathcal{F}|$
Effective cooperation weight $\kappa_i \leftarrow \tau_i c_i$
Reservation utilities $d_i, d_j \leftarrow 0$
for $t = 1$ to $T$ do
(a) Compute trust–confidence–gated moral utilities:
    for each joint action $(a_i, a_j)$ do
        $U_i(a_i, a_j) = \sum_{\phi\in\mathcal{F}} \big[(1-\kappa_i)\, B_i(\phi) + \kappa_i\, \hat{B}_{i\to j}(\phi)\big] F_\phi(a_i, a_j)$
        $U_j(a_j, a_i) = \sum_{\phi\in\mathcal{F}} \big[(1-\kappa_j)\, B_j(\phi) + \kappa_j\, \hat{B}_{j\to i}(\phi)\big] F_\phi(a_j, a_i)$
(b) Joint Virtual Bargaining (conceptual coordination step):    ▷ Both agents' utilities are evaluated to determine the counterfactual joint action.
    $(a_i^{(\mathrm{VB})}, a_j^{(\mathrm{VB})}) = \arg\max_{a_i, a_j} \big[(U_i(a_i,a_j) - d_i)_+\big]^{\gamma} \big[(U_j(a_j,a_i) - d_j)_+\big]^{1-\gamma}$
        ▷ $(x)_+ = \max(x, 0)$; $\gamma \in (0,1)$ balances bargaining asymmetry
(c) Fairness deviation and trust update:
    Define fairness-only utility under the agent's own moral prior:
        $U_i^F(a_i, a_j) = \sum_{\phi\in\mathcal{F}} B_i(\phi)\, F_\phi(a_i, a_j)$
    Compute fairness deviation as the gap between ideal and realized fairness outcomes:
        $d_i^{(t)} = \max_{a'_j \in \mathcal{A}_j} U_i^F(a_i^{(\mathrm{VB})}, a'_j) - U_i^F(a_i^{(\mathrm{VB})}, a_j^{(\mathrm{VB})})$
        ▷ $d_i^{(t)}$ measures the fairness shortfall caused by partner $j$'s action.
    Compute compliance $s_i^{(t)} = \exp(-\lambda_{\mathrm{dev}} d_i^{(t)})$
    Update trust:
        $\tau_i^{(t+1)} = (1-\eta)\tau_i^{(t)} + \eta\, s_i^{(t)}$
(d) Belief update (CK-ToM inference):
    for each fairness norm $\phi \in \mathcal{F}$ do
        $\hat{B}_{i\to j}^{(t+1)}(\phi) \propto \hat{B}_{i\to j}^{(t)}(\phi) \exp\big[\beta F_\phi(a_i^{(\mathrm{VB})}, a_j^{(\mathrm{VB})})\big]^{1-\alpha\tau_i^{(t)}} [B_i(\phi)]^{\alpha\tau_i^{(t)}}$
    Normalize beliefs: $\sum_\phi \hat{B}_{i\to j}^{(t+1)}(\phi) = 1$
    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 utility: $d_i^{(t+1)} = U_i^F(a_i^{(\mathrm{VB})}, a_j^{(\mathrm{VB})})$
```

noise term ϵ_p applied to fairness evaluations $F_\phi(a_i, a_j)$. After each interaction, a new (R_i, F_ϕ) pair is resampled, producing an open-ended stream of moral dilemmas that vary jointly in material asymmetry and normative structure. All payoff and fairness values are rescaled to $[0, 1]$, and the parameter grid $(\epsilon_a, \epsilon_p, T, s, \mu)$ follows the settings detailed in the *Supplementary Appendix*. MGG generation code, random seeds, and configuration files used in all simulations are included in the repository to ensure full reproducibility.

Baseline and Comparison Agents. To benchmark TGMI, we evaluated it alongside established reciprocal and automata strategies from evolutionary game theory. These include **Tit-for-Tat (TFT)**, **Generous Tit-for-Tat (GTFT)**, **Win-Stay–Lose-Shift (WSLS)**, **Forgiver**, **Always Cooperate (AllC)**, **Always Defect (AllD)**, and **Zero-Determinant Extortion (ZD-Extort)** agents. Together, these baselines span unconditional cooperation, strict reciprocity, generosity, and exploitation. Behavioral rules follow canonical definitions in (**? ?**): AllC, TFT, WSLS, GTFT, and Forgiver begin by cooperating, while AllD and ZD-Extort defect initially.

For completeness, we also evaluated three **TGMI ablations** to isolate the model's core mechanisms: (1) *no-trust*, which removes trust updates and treats all interactions as morally neutral; (2) *no-belief*, which fixes moral priors without inference; and (3) *no-bargaining*, which omits the virtual bargaining step, reverting to single-agent moral optimization. These variants clarify the distinct contributions of trust gating, moral inference, and counterfactual coordination to cooperative stability.

Evolutionary Analysis of the Moral Game Generator. We study the long-run abundance of strategies under **finite-population social learning** using a Moran process. The population size is $N = 10$, mutation rate $\mu = 10^{-3}$, and selection strength $s = 2$. Each generation, agents interact in the **Moral Game Generator (MGG)** for T rounds with public observability q (the fraction of interactions observed by all), action error ϵ_a, and optional perception noise ϵ_p.

Let

$$\mathcal{T} = \{\text{TGMI, Selfish, Altruistic, TFT, GTFT, WSLS, Forgiver, ZD}\}$$

denote the set of agent types (including TGMI ablations when applicable). A population composition is represented by $\mathbf{n} = (n_1, \ldots, n_M)$ with $\sum_k n_k = N$, where $M = |\mathcal{T}|$. The state space is the set of all such $\mathbf{n}$.

For a focal agent i of type k in composition $\mathbf{n}$, let $R_i^{(t)}$ denote the material payoff at round t, and $U_i^{F,t}$ the fairness utility (Eq. 2, main text). We define a **fairness-weighted return** with mixing parameter $\omega \in [0, 1]$:

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

This allows evolutionary selection to depend jointly on material and moral satisfaction, enabling comparison between value-aligned and purely strategic agents ($\omega = 0$ for payoff-only selection; $\omega = 0.25$ in robustness checks).

Expected returns for each type k are estimated by Monte Carlo averaging over $S = 200$ replicates:

$$\bar{\Phi}_k(\mathbf{n}) = \frac{1}{S}\sum_{s=1}^{S}\Phi_{i,s}(\mathbf{n}), \quad \text{for all } k.$$

Expected returns are then mapped to **copying probabilities** via a softmax:

$$\pi_k(\mathbf{n}) = \frac{\exp(s\bar{\Phi}_k(\mathbf{n}))}{\sum_{m=1}^{M}\exp(s\bar{\Phi}_m(\mathbf{n}))}.$$

At each generation, a random learner copies type k with probability $(1-\mu)\pi_k(\mathbf{n})$ or mutates uniformly with probability μ/M. The one-step transition probability between population compositions is

$$P(\mathbf{n} \to \mathbf{n} + e_k - e_\ell) = \left[(1-\mu)\pi_k(\mathbf{n}) + \frac{\mu}{M}\right]\frac{n_\ell}{N}.$$

Collecting all transitions yields a stochastic matrix P, whose stationary distribution x^* satisfies $x^* = x^* P$. We compute x^* as the dominant left eigenvector of P (normalized to sum to 1) using power iteration until convergence ($\|x^{(t+1)} - x^{(t)}\| < 10^{-9}$). The steady-state abundance of type k is given by

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

We report $\bar{x}_k^*$ and population-level welfare and fairness statistics across $(T, q, \epsilon_a, \epsilon_p, \omega)$, with robustness checks over s and μ.

We use two-player MGG episodes for intragenerational returns and two–three-player episodes for evolutionary runs. Actions are continuous; TGMI beliefs are initialized $\hat{B}_{i\to j} \sim \text{Dirichlet}(1, \ldots, 1)$; moral priors $B_i(\phi)$ are Dirichlet-sampled over $\mathcal{F}$; trust τ_i starts at $\tau_0 \in [0, 1]$. Public observability q determines whether CK-ToM updates are applied globally or dyadically. Parameter grids and random seeds are provided in the *Supplementary Appendix.*

Control comparisons. To connect TGMI with canonical reciprocity results, we include a *payoff-weighted control* replacing moral fallback with payoff fallback:

$$U_i^{\text{ctrl}}(a_i, a_j) = (1-\tau_i)R_i(a_i, a_j) + \tau_i R_j(a_i, a_j),$$

in a two-player IPD with $R_i(a_i, a_j) = Ba_j - Ca_i$ $(B > C > 0)$. This control reproduces the classical Bayesian Reciprocator when fairness reduces to a single efficiency norm, $\mathcal{F} = \{\text{Max-Sum}\}$, where

$$F_{\text{Max-Sum}}(a_i, a_j) = R_i(a_i, a_j) + R_j(a_i, a_j).$$

Under this reduction, TGMI subsumes payoff-based reciprocity as a special case, demonstrating that moral inference generalizes classical cooperation rather than replacing it.

Simulation details and steady-state abundance results are reported in Fig. Sx and Table Sy of the *Supplementary Appendix.*

Data, Materials, and Software Availability. Simulation code, parameter settings, and analysis scripts are available at `[GitHub Repository Placeholder]`.

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.