Why human societies adopt rigid moral rules: the efficiency–robustness trade-off

Julien Lie-Panis[a,1], Léo Fitouchi[b], Nicolas Baumard[c], and Jean-Baptiste André[c]

This manuscript was compiled on March 23, 2026

Humans are capable of remarkably flexible moral judgment. Yet societies rely on rigid rules—obligations and prohibitions that apply categorically, even when case-by-case reasoning could yield better outcomes. Why would a species capable of flexibility bind itself to rigid rules? We propose that rigid rules arise as social technologies for managing ambiguity about noncooperation. People often have legitimate reasons for failing to cooperate, yet those reasons are typically opaque to observers, allowing opportunists to disguise selfishness as justified hardship. We formalize this idea with an evolutionary game-theoretic model. Two cooperative equilibria emerge: a flexible norm that accommodates legitimate excuses but is vulnerable to exploitation, and a rigid norm that closes this loophole by mandating cooperation even when inefficient. Comparing these equilibria reveals an efficiency–robustness trade-off: flexibility maximizes welfare when trust is secure, whereas rigidity preserves cooperation when trust is fragile. This explains why rigid rules prevail in low-trust settings—interactions with strangers, formal institutions, or tight societies—while flexibility is more common in high-trust contexts.

morality | deontology | cooperation | norms | evolutionary game theory

"Please keep off the grass—unless you have a good reason." This sign, once posted on the lawn of an Oxford college, represents a flexible approach to collective action. Rather than enforcing a categorical prohibition ("Keep off the grass!"), it invited individuals to exercise judgment, distinguishing between mere convenience (avoiding a short detour) and genuine need (rushing to the hospital). Rigidly staying off the grass would surely keep the lawn immaculate, but at needless cost. By allowing people to weigh the costs and benefits of every detour, the rule achieved its goal efficiently, protecting the grass while permitting legitimate exceptions.

Humans routinely exercise this kind of context-sensitive moral reasoning, adjusting their judgments to the fine-grained parameters of each situation (1–11). When waiting in line, we readily distinguish between someone who cuts out of impatience and someone who cuts because they are diabetic and urgently need sugar (12). This capacity for flexible moral judgment emerges by middle childhood (13–15), suggesting that it is a central feature of human cognition.

Yet few social rules make use of this human flexibility. Across societies—from small-scale foraging bands (16, 17) to large nation-states (18)—cooperation rests on standardized rules that hold regardless of context. In industrialized societies, drivers must obey speed limits and stop at red lights even when roads are empty and conditions perfectly safe. Laws prohibit stealing others' property regardless of need, although people readily judge theft to avert starvation as more excusable than theft for greed (19–21; cf. *Southwark London Borough Council v Williams* 22). In many subsistence societies, elaborate sharing rules divide a kill by fixed roles—the spotter, the first striker, the harpoon owner— irrespective of who contributed most that day (23–27). Across these settings, rules apply categorically, even when taking context into account could make cooperation more efficient.

Significance

Human moral life is governed by rigid rules. Across societies, people rely on categorical prohibitions and duties—do not lie, do not steal, always keep your promises—even when they recognize compelling reasons to make exceptions. Using a mathematical model, we show that rigid rules can protect cooperation when motives are opaque. Flexible rules allow justified exceptions but also make it easier for cheaters to hide behind plausible excuses. This creates a trade-off between maximizing welfare and protecting cooperation from those who would exploit excuses to justify selfishness. Our account explains why rigid rules are common in low-trust settings, such as with strangers or in rule-bound societies, and why many moral judgments feel like non-negotiable duties rather than cost-benefit calculations.

This contrast between flexible minds and rigid rules presents a puzzle: why would a species of flexible cooperators bind itself to categorical rules that demand cooperation even when it is inefficient?

We propose that rigid norms persist because they solve a problem intrinsic to reputation-based cooperation: opacity about why people fail to cooperate.

To illustrate, imagine two roommates, Alice and Bob, trying to divide the household chores. They could adopt a flexible arrangement, letting each clean when they judge, in good faith, that it's their turn. This promises efficiency: there's no point in requiring someone to scrub the bathroom when they're sick or buried in work—the cost to them outweighs the benefit of a cleaner apartment.

But flexibility creates moral wiggle room: ambiguity that an opportunist can exploit (e.g., 28, 29). If Bob walks into a cluttered kitchen and Alice reports being overwhelmed with work, he cannot tell whether this reflects genuine hardship or a convenient excuse. That opacity shifts Alice's incentives by giving her cover to exaggerate her workload when she feels lazy. Flexibility thus opens the door to strategic defection, leaving opportunists room to exploit reputational ambiguity.

A rigid rule closes this loophole. Suppose chores are assigned on a fixed schedule—Alice cleans on Sundays and Bob on Thursdays. Failing to clean on one's scheduled day is then unambiguously a defection. The rigid rule draws a bright line between cooperation and defection, eliminating the ambiguity that opportunists depend on.

We formalize this idea with a model of reputation-based cooperation. In the model, individuals sometimes face genuine hardship—illness, heavy workload, or personal crises—that makes cooperation socially inefficient. Because observers cannot reliably verify these circumstances, actors have room to disguise selfish behavior as justified hardship.

We identify two distinct cooperative equilibria. Under a flexible norm, cooperation accommodates context: individuals who appear to face hardship retain others' trust. Under a rigid norm, by contrast, cooperation is demanded categorically—any failure to cooperate leads to loss of trust, regardless of apparent circumstances. Comparing these equilibria reveals a fundamental trade-off between efficiency and robustness. Flexibility is efficient, allowing individuals to opt out when cooperation would be genuinely costly, but it leaves observers vulnerable to strategic defection. Rigidity, though occasionally inefficient, makes cooperation more robust, extending it to environments where trust is fragile and flexibility too risky.

In this sense, rigid moral rules act as a social technology: they trade efficiency for robustness, extending the reach of reputation-based cooperation. Based on this model, we predict that flexibility will prevail in high trust environments—such as families, friendships, and tightly knit teams—or when dealing with easily verified hardships like physical injury. Rigid rules, by contrast, should emerge among strangers, in large groups, or in low-trust societies.

More broadly, this logic helps explain a central puzzle in moral psychology: why humans often treat certain rules and values as sacred, insisting on duty regardless of consequence. What appears as deontological rigidity may instead reflect an adaptive response to the problem of sustaining cooperation under uncertainty.

Model

We develop a model of reputation-based cooperation where individuals sometimes face genuine hardship—situations in which cooperating would be too costly—but others cannot tell for sure whether this is the case.

Players and basic structure. We consider a repeated game with two roles: $n \gg 1$ signalers and an infinite pool of choosers. Signalers make cooperative decisions across rounds and hold reputations that reflect their actions. Choosers, in contrast, participate in only one round and decide whether to trust a signaler based on their reputation.

This setup separates two adaptive problems. For choosers, the challenge is trust—predicting whether a partner will reciprocate trust if given the opportunity. For signalers, the challenge is reputation management—acting in ways that secure the trust of future partners.

To capture variation in reputation management strategies, we assume signalers differ in their time preferences. Each signaler is characterized by a discount factor δ (with values between 0 and 1), which measures how much they value future payoffs relative to immediate ones. More patient signalers (higher δ) are more willing to incur the immediate costs of cooperation to maintain their reputation with choosers, whereas impatient individuals (lower δ) prioritize short-term gains and are more likely to defect.

We assume that discount factors are continuously distributed across the population. This produces a spectrum of patience levels, capturing the full range of cooperative tendencies. In particular, it allows for strategic actors who exploit flexible norms but are restrained by rigid ones—the kind of behavior that rigid rules are designed to prevent.

As illustrated in Figure 1, the game proceeds in two phases: signalers first participate in a public goods game, before interacting with choosers in trust games.

Initial collective action under potential hardship. In the initial interaction, signalers decide whether to contribute to a collective action or free-ride on others' contributions. They play a linear public goods game, where each contribution imposes a private cost on the signaler but generates a public benefit for others.

The cost of contribution depends on circumstances. Under normal conditions, contribution imposes a relatively low private cost γ. However, with probability p, a signaler experiences hardship—illness, competing obligations, or a personal crisis—that increases their cost of contribution to γ_H.

Each contribution yields a public benefit β, which is shared equally among the other signalers (excluding the contributor). The net social value of a contribution is therefore $\beta - \gamma$ under normal conditions and $\beta - \gamma_H$ under hardship. We assume:
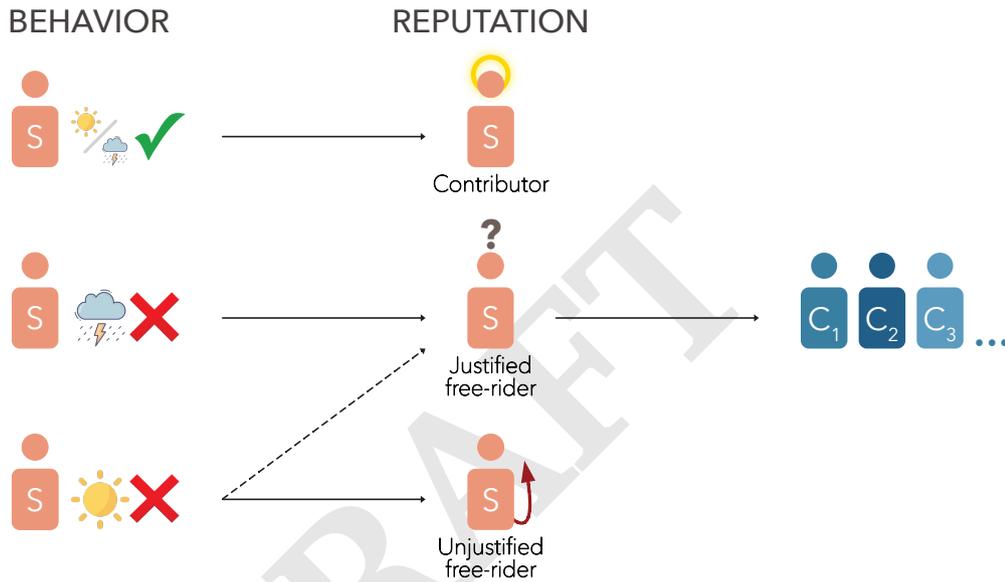
$$0 < \gamma < \beta < \gamma_H,$$

making contributions socially efficient under normal conditions ($\beta - \gamma > 0$), but inefficient under hardship ($\beta - \gamma_H < 0$).

251
252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
315
316
317
318
319
320
321
322
323
324
325
326
327
328
329
330

a Initial collective action
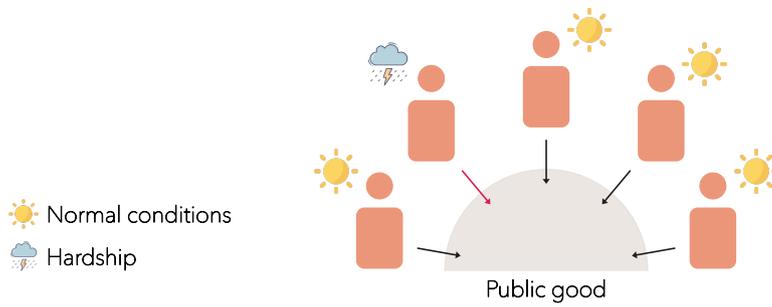
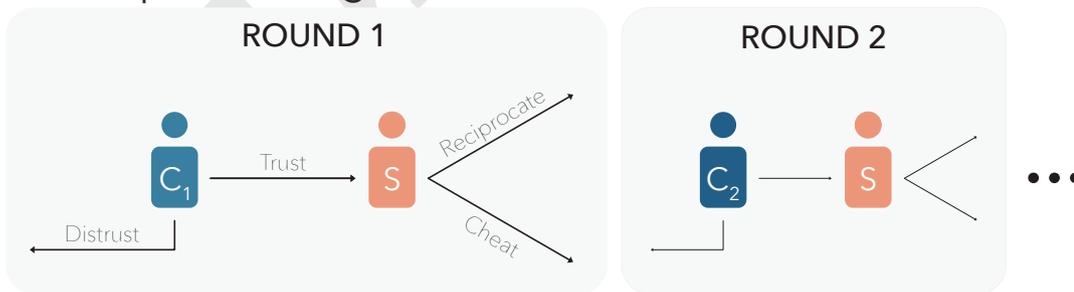## b Reputation formation

c Subsequent trust games

Fig. 1 — Model overview. a, Initially, signalers decide whether to contribute to a public good. Contribution is efficient under normal conditions but inefficient under hardship. **b,** Signalers acquire reputations based on their behavior. Those who free-ride under normal conditions are labeled justified free-riders with probability ε, making them indistinguishable from those who free-ride under genuine hardship. The parameter ε captures plausible deniability—how easily opportunists can be mistaken for genuinely hard-pressed individuals. **c,** Subsequently, signalers interact with choosers, who decide whether to trust based on reputation; if trusted, signalers decide whether to reciprocate.

Signalers decide whether to contribute based on their circumstances. Choosers perfectly observe the outcome (contribution or free-riding), but not the underlying circumstances that led to that decision.

As shown in Figure 1b, this partial observability generates reputational ambiguity. Signalers who contribute become contributors, regardless of circumstances. Those who free-ride under hardship become justified free-riders. However, those who free-ride under normal conditions

mistakenly become justified free-riders with probability ε, and are otherwise labeled unjustified free-riders.

The parameter ε captures the degree of plausible deniability: how easily opportunistic free-riders can appear genuinely hard-pressed. We model this as observational noise for simplicity, but an equivalent formulation would let individuals who free-ride under normal conditions falsely claim hardship at no cost, succeeding with probability ε—an opportunity they would all take in equilibrium.

This assumption preserves the strategic logic of plausible deniability without adding unnecessary complexity.

Subsequent trust games. After the initial stage, each signaler interacts with a sequence of choosers. In every round, a new chooser decides whether to trust the signaler, incurring cost c to provide benefit b (with $0 < c < b$). If trusted, the signaler can reciprocate—incurring cost c to return benefit b to the chooser—or cheat, keeping the benefit without giving back. These trust games are one-shot for choosers but repeated for signalers, creating an incentive to maintain a cooperative reputation over time.

After each round, a signaler's reputation updates to reflect their most recent behavior. The full reputation combines their initial collective action label (contributor, justified free-rider, or unjustified free-rider) with their most recent behavior in trust games (reciprocator or cheater), if any. Choosers rely on this composite reputation when deciding whether to place their trust.

Results

The analysis reveals two cooperative equilibria, Perfect Bayesian Equilibria that sustain reputation-based cooperation (see Methods for a formal definition): a flexible norm that accommodates apparent hardship, and a rigid norm that categorically demands cooperation regardless of context.

For clarity, we assume:

$$\frac{\gamma}{1 - \varepsilon} < c < \gamma_H.$$

This assumption ensures that contributing under normal conditions is less demanding than reciprocating trust (left inequality), and that contributing under hardship is more demanding (right inequality). It simplifies the behavioral thresholds below without changing any qualitative result (see Supplementary Information for the general case).

Two equilibria: flexible and rigid norms.

Flexible norm. In this equilibrium, choosers interpret free-riding leniently: if a signaler appears to have faced hardship, they are still deemed trustworthy. In the initial trust game, before a signaler's reciprocation behavior is known, trust is extended to contributors and justified free-riders, but withheld from unjustified free-riders. Later, once reciprocation is observed, choosers trust reciprocators and distrust cheaters.

Signalers never contribute under hardship, since free-riding then carries no reputational penalty. Under normal circumstances, however, they face a trade-off: free-riding avoids the immediate cost of contribution, γ, but risks losing the next chooser's trust—worth b—with probability $1 - \varepsilon$, if the signaler's true circumstances are observed. Signalers then contribute if and only if their discount factor satisfies $\delta \geq \delta_{\text{normal}}^{\text{flex.}}$, where:

$$\delta_{\text{normal}}^{\text{flex.}} = \frac{\gamma}{(1 - \varepsilon)b}. \qquad \textbf{[F1]}$$

In subsequent trust games, reciprocation incurs cost c but secures next chooser's trust, worth b. Signalers reciprocate if and only if their discount factor satisfies $\delta \geq \delta_{\text{recip.}}$, where:

$$\delta_{\text{recip.}} = \frac{c}{b}.$$

Because $\frac{\gamma}{1 - \varepsilon} < c$, this threshold is higher than $\delta_{\text{normal}}^{\text{flex.}}$; contributing under normal circumstances requires less patience than reciprocating trust.

From these thresholds, three signaler types emerge. Impatient signalers ($\delta < \delta_{\text{normal}}^{\text{flex.}}$) always free-ride and always cheat, regardless of circumstances or reputation. Intermediate signalers ($\delta_{\text{normal}}^{\text{flex.}} \leq \delta < \delta_{\text{recip.}}$) contribute under normal circumstances but cheat when trusted. Patient signalers ($\delta \geq \delta_{\text{recip.}}$) contribute under normal circumstances and always reciprocate trust.

The flexible norm is sustainable whenever initial trust decisions are well-founded. The binding constraint involves justified free-riders, who must reciprocate frequently enough to make trusting them worthwhile. The flexible norm exists if and only if:

$$\frac{p\,\mathbb{P}(\delta \geq \delta_{\text{recip.}})}{p + (1 - p)\varepsilon\,\mathbb{P}(\delta < \delta_{\text{normal}}^{\text{flex.}})} \geq \frac{c}{b}. \qquad \textbf{[F2]}$$

This condition ensures that justified free-riders are, on average, sufficiently likely to reciprocate relative to the cost-to-benefit ratio of trust, $\frac{c}{b}$. The numerator gives the probability that a reciprocator acquires the justified free-rider reputation—necessarily by facing hardship—while the denominator gives the overall probability of that reputation: all signalers become justified free-riders when they face hardship, with probability p, and free-riders also acquire this reputation when they face normal conditions and successfully misrepresent them as hardship, with probability $(1 - p)\varepsilon$.

Rigid norm. In this equilibrium, choosers adopt a zero-tolerance policy: only contributors are trusted, while any instance of free-riding—regardless of apparent circumstances—results in loss of trust. In the initial trust game, only contributors are trusted, while both justified and unjustified free-riders are distrusted. Later, choosers trust reciprocators and distrust cheaters.

Signalers now face stronger incentives to contribute under normal conditions. They do so if and only if their discount factor satisfies $\delta \geq \delta_{\text{normal}}^{\text{rigid}}$, where:

$$\delta_{\text{normal}}^{\text{rigid}} = \frac{\gamma}{b}. \qquad \textbf{[R1]}$$

This threshold is lower than under the flexible norm ($\frac{\gamma}{b} < \frac{\gamma}{b(1 - \varepsilon)}$). Signalers whose discount factor falls between these two thresholds can be understood as opportunists: they contribute when rigidity prevents misrepresentation but free-ride when flexibility allows reputational ambiguity to be exploited—opportunists, in short, exploit moral wiggle room when it is available.

The rigid norm also creates incentives to contribute under hardship. A signaler facing hardship contributes if and only if their discount factor satisfies $\delta \geq \delta_{\text{hardship}}^{\text{rigid}}$, where:

$$\delta_{\text{hardship}}^{\text{rigid}} = \frac{\gamma_H}{\gamma_H + b - c}. \qquad \textbf{[R2]}$$

Signalers reciprocate trust if and only if their discount factor satisfies $\delta \geq \delta_{\text{recip.}} = \frac{c}{b}$, as with the flexible
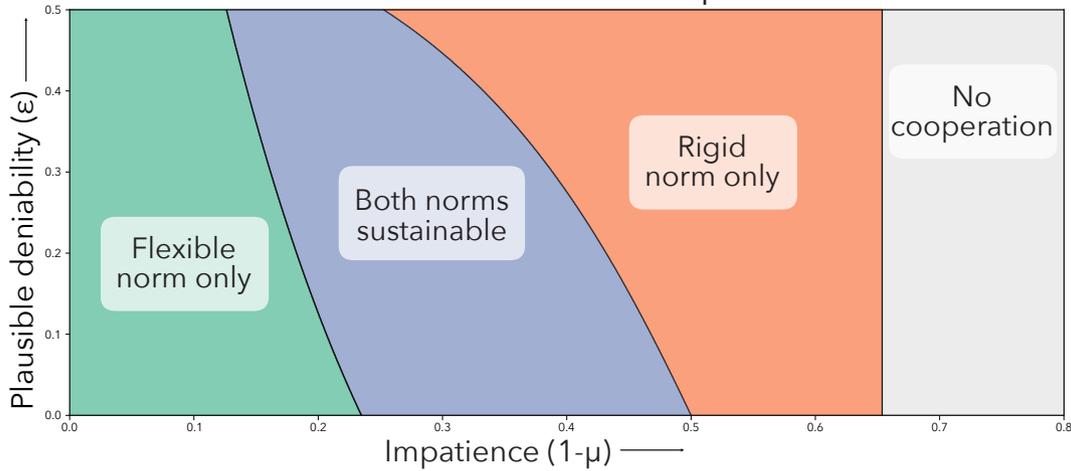
Domains of sustainable cooperation

Fig. 2 — Domains of sustainable cooperation. Each point represents a social environment defined by impatience $1 - \mu$ (x-axis) and plausible deniability ε (y-axis). Colors indicate where the flexible (F2) and rigid (R3-R4) norms are sustainable. Green: only the flexible norm is sustainable; orange: only the rigid norm; blue: both; white: cooperation fails. Discount factors follow a normal distribution with mean μ and standard deviation $\sigma = 0.3$, truncated to the interval $(0, 1)$. Parameters: $c = 1, b = 2, \gamma = 0.5, \beta = 0.75, \gamma_H = 2, p = 0.2$. We vary impatience $1 - \mu$ between 0 and 0.75 (i.e., μ between 1 and 0.25) and plausible deniability ε between 0 and $1 - \frac{\gamma}{c} = 0.5$, ensuring that $\gamma/(1 - \varepsilon) < c$ holds throughout.

norm. From these thresholds, four types of signalers emerge. Very impatient signalers ($\delta < \delta_{\text{normal}}^{\text{rigid}}$) always free-ride and cheat; impatient signalers ($\delta_{\text{normal}}^{\text{rigid}} \leq \delta < \delta_{\text{recip.}}$) contribute under normal circumstances but cheat when trusted; patient signalers ($\delta_{\text{recip.}} \leq \delta < \delta_{\text{hardship}}^{\text{rigid}}$) contribute under normal circumstances and reciprocate; and very patient signalers ($\delta \geq \delta_{\text{hardship}}^{\text{rigid}}$) contribute even under hardship and reciprocate.

The rigid norm is sustainable whenever initial trust decisions are well-founded. There are two binding conditions, involving contributors and justified free-riders respectively. The rigid norm exists if and only if:

$$\frac{p \, \mathbb{P}(\delta_{\text{recip.}} \leq \delta < \delta_{\text{hardship}}^{\text{rigid}})}{p \, \mathbb{P}(\delta < \delta_{\text{hardship}}^{\text{rigid}}) + (1-p)\varepsilon \, \mathbb{P}(\delta < \delta_{\text{normal}}^{\text{rigid}})} < \frac{c}{b}, \quad \textbf{[R3]}$$

$$\frac{(1-p) \, \mathbb{P}(\delta \geq \delta_{\text{recip.}}) + p \, \mathbb{P}(\delta \geq \delta_{\text{hardship}}^{\text{rigid}})}{(1-p) \, \mathbb{P}(\delta \geq \delta_{\text{normal}}^{\text{rigid}}) + p \, \mathbb{P}(\delta \geq \delta_{\text{hardship}}^{\text{rigid}})} \geq \frac{c}{b}. \quad \textbf{[R4]}$$

Condition (R3) ensures that justified free-riders are sufficiently unlikely to reciprocate, making it sensible to distrust them, while condition (R4) ensures that contributors are sufficiently likely to reciprocate, making trust worthwhile. In both cases, the numerators give the probability that a reciprocator holds the relevant reputation, and the denominators give its overall probability in the population.

Comparing flexible and rigid norms. These two equilibria represent distinct approaches to moral judgment under uncertainty. The flexible norm extends trust to those who appear to have faced hardship; the rigid norm withholds it. As captured by conditions (F2) and (R3), whether such trust is well-founded depends on how likely these justified free-riders are to reciprocate trust.

That likelihood, in turn, depends on two features of the environment: the distribution of discount factors, which shapes the population's cooperativeness, and plausible deniability, which measures how easily selfish behavior can be disguised as hardship.

To illustrate, we assume signalers' discount factors follow a normal distribution with mean μ and standard deviation σ, truncated to the interval $(0, 1)$. Figure 2 shows the resulting domains of sustainable cooperation as functions of impatience $(1 - \mu)$ and plausible deniability (ε). When the population is patient and deniability is low, only the flexible norm is sustainable: trust can safely extend to justified free-riders, who are likely to reciprocate and to have faced genuine hardship. As impatience and deniability increase, the flexible norm collapses, and cooperation can persist only under the rigid norm. When impatience becomes extreme, even contributors can no longer be trusted, and cooperation breaks down altogether.

Rigid norms do more than offer an alternative to flexible ones—they enable cooperation in contexts where flexibility would fail. Figure 3 shows how rigid norms achieve this by comparing performance where both norms are possible. The rigid norm is more robust to strategic defection but less efficient: it prevents opportunistic exploitation of ambiguity (Figure 3a), but incentivizes cooperation under hardship, leading to an overall payoff loss (Figure 3b).

Thus, rigid norms function as social technologies that trade efficiency for robustness against strategic defection, extending the domain of cooperation by eliminating exploitable ambiguity.

Discussion

Rigid rules as social technologies. Why do societies adopt rigid moral rules when humans are perfectly capable of flexible, context-sensitive moral reasoning? Our model addresses this question by examining how cooperation evolves when people's circumstances are private and reputations ambiguous.

Individuals in our model sometimes face genuine hardship—situations in which cooperating would be socially inefficient. Yet observers cannot perfectly assess
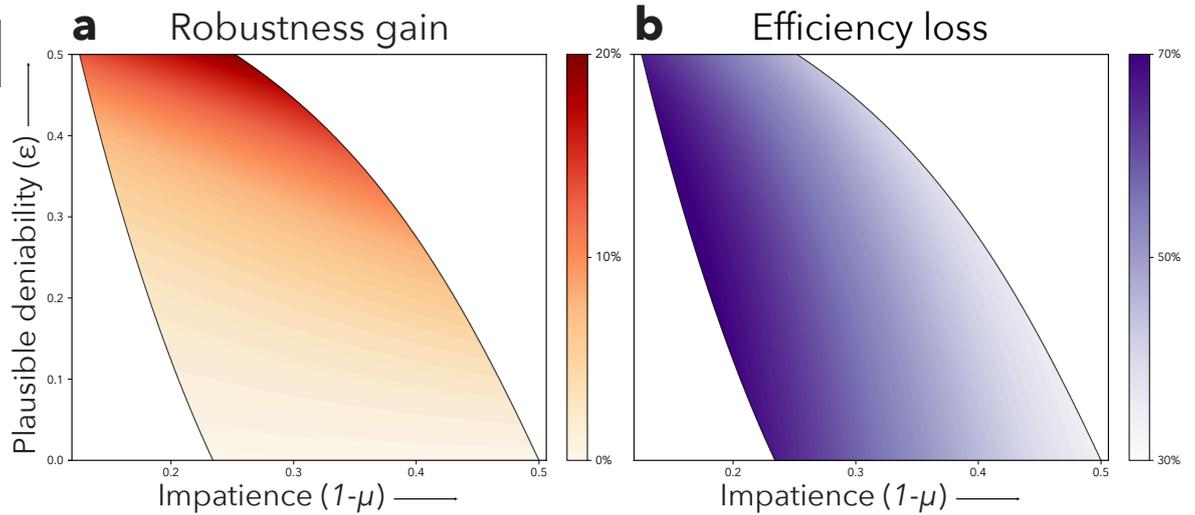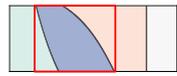
a Robustness gain   **b** Efficiency loss

Fig. 3 — The efficiency–robustness trade-off. Each point represents a social environment defined by impatience $1 - \mu$ (x-axis) and plausible deniability ε (y-axis). We restrict attention to the domain where both norms are sustainable (blue region in Figure 2) and compare their performance. **a,** Robustness gain from adopting the rigid norm. Warm shades (yellow to red) show the fraction of opportunists who, under normal conditions, would free-ride under the flexible norm but contribute when cooperation is categorically required. **b,** Efficiency loss from adopting the rigid norm. Purple shades show the normalized payoff difference from the initial collective action (flexible – rigid). Parameters and distribution of discount factors as in Figure 2.

these circumstances: with some probability, an individual who defects under normal conditions is mistaken for someone acting under hardship. This probability captures the degree of plausible deniability—how easily selfish behavior can appear as genuine constraint. Individuals also differ in their time preferences: more future-oriented individuals are more willing to incur the immediate costs of cooperation to maintain their reputation, whereas present-oriented individuals prioritize short-term gains.

The analysis reveals two cooperative equilibria, representing two approaches to moral judgment under uncertainty: a flexible norm that accommodates apparent hardship, and a rigid norm that categorically demands cooperation regardless of context. These equilibria succeed under different conditions. The flexible norm is sustainable when trust is secure; when plausible deniability is low and individuals are future-oriented, those who appear hard-pressed are genuinely trustworthy, and contextual discretion does not invite exploitation. In contrast, when individuals are present-oriented and hardship is easy to fake, flexible trust becomes too risky, but cooperation can be sustained under the rigid norm.

The rigid norm thus extends cooperation to settings where flexible trust would fail. Comparing it to the flexible norm reveals an efficiency–robustness trade-off: by treating all defections alike, rigidity makes trust decisions robust to strategic defection, but at the cost of sometimes enforcing inefficient cooperation. Rigid norms appear as social technologies that trade efficiency for robustness.

Empirical predictions. Our model predicts that rigid rules—those that demand cooperation even in apparent hardship—should arise when flexibility is most risky; that is, when the potential for strategic defection is highest.

First, rigidity should emerge around circumstances that are easy to fake. Consider pregnancy: women in late stages, where their condition is more visible, often receive accommodation, such as seats on public transport or flexibility at work. Such accommodations are much rarer in early pregnancy, when needs are comparable but

less apparent. More generally, hard-to-verify claims like stress or fatigue should receive less accommodation than observable injuries.

Second, moral rules should be applied more rigidly to people perceived as impulsive or untrustworthy. This may explain why young children are taught categorical moral rules with little room for contextual nuance—they are not yet seen as reliable enough to exercise moral discretion (see also 13). A similar logic shapes attitudes toward welfare: when recipients are perceived as lazy rather than unlucky, people tend to favor strict eligibility rules (30).

Third, rigid rules should be more frequent in distant relationships. We predict that individuals should cooperate more flexibly with family and friends, but adopt more rigid rules with colleagues or strangers. This sheds light on relational psychology (31). Recent work suggests that formality is a fundamental dimension of relational psychology, distinguishing public relationships that adhere to rules and regulations from private relationships with looser, more casual styles (32). Consistent with our prediction, relationships become more formal in modernized societies, where interactions increasingly involve strangers rather than intimate social circles.

Fourth, distrust should foster rigidity. This helps explain why some societies are tight—with many strong norms and a low tolerance of deviant behavior—and others are loose (18). Tightness is associated with ecological and social threats, such as disease, natural disasters, resource scarcity, population density, and intergroup conflict (33–35). Existing accounts attribute this pattern to coordination demands: groups under threat require stronger norms to sustain collective action (formalized in 36). Our model complements this view by identifying a cognitive mechanism that makes rigidity appealing. In threatening environments, immediate needs loom larger (37), making investment in one's cooperative reputation less worthwhile. Recognizing this, people expect others to be less reliably cooperative and see flexibility as too risky, turning instead to more rigid rules—the loss in efficiency is outweighed by the gain in robustness. As

shown by Nettle and Saxe (38), a similar mechanism underlies support for authoritarian leaders and punitive regulation, explaining why these political preferences also rise under threat (39–42).

The logic of moral rigidity. Moral judgments are often strikingly deontological: people treat certain rules, values, and rights as sacred or inalienable, refusing to weigh them against costs, benefits, or pragmatic compromises (1, 43–46). One way to understand this rigidity is through the efficiency–robustness trade-off. As we have shown, categorical rules close loopholes, making cooperation easier to sustain. The same logic extends more broadly: rigid principles keep expectations clear, make departures easier to detect or sanction, and prevent minor allowances from hardening into precedents that legitimize further slippage, gradually weakening the cooperative order (47). People may therefore endorse some rules categorically not because they view them as intrinsically right, but because they sense that strict adherence is the most reliable way to sustain cooperation in the long run.

Our account complements existing explanations for moral rigidity. Rules can act as cognitive shortcuts (48), helping individuals approximate—and coordinate on—mutually beneficial bargains (49). Bright-line norms can also facilitate coordination among third-party enforcers (50) and bystanders (51). And principled, trade-off–insensitive behavior can serve as a signal of commitment, helping individuals build trust (52, 53).

Implications for evolutionary game theory. Our model differs from the standard framework for reputation-based cooperation in three ways.

First, we introduce heterogeneity in cooperative types. In standard indirect reciprocity models, individuals are identical and interactions are repeated (54–56). Norms—understood as rules for assigning good or bad reputation—are then evaluated by their capacity to generate and stabilize cooperation (57–61). We instead assume that signalers vary in patience, integrating the logic of signaling with that of repeated cooperation (building on 62; see also 63, 64). Patient signalers reciprocate trust to attract the trust of future partners, whereas impatient ones cheat. Different norms then emerge depending on which reputations predict reciprocation and which predict cheating. This avoids the scoring dilemma: cooperation in the trust game is sustainable because first-order reputations (reciprocator or cheater) perfectly predict future behavior.

Second, we add a public goods game, following the structure of Panchanathan and Boyd (65). The signaling framework clarifies when contributions can be stabilized: signalers must be sufficiently patient for choosers to infer trustworthiness from contribution.

Third, we introduce opaque exculpatory circumstances, a feature rarely modeled (although see 66). This creates opportunities for strategic misrepresentation, leading to reputational ambiguity. Flexible assessments emerge when choosers can trust despite this ambiguity; rigidity becomes necessary when trust is too risky. Adopting a signaling perspective thus reframes norm evolution: stricter assessment rules—akin to Stern Judging in indirect reciprocity models (61, 67–69)—arise as adaptive responses to observers' inferential challenges.

Limits and future directions. Our model isolates one adaptive problem—judging trustworthiness under opaque circumstances—and shows how it can produce both flexible and rigid assessments. But moral norms are not just solutions to individual inference problems. They involve widely shared expectations about how people ought to behave, the felt obligation to conform, and culturally inherited conventions that come to seem external or binding (70). Real-world cooperation also requires forms of coordination that our analysis sets aside—coordination on the standards we use to assess others, on when and how to enforce them, and on the shared courses of action that make collective endeavors possible. Integrating our approach with models that have examined such dynamics (71–74) is a natural direction for future work.

Conclusion

Flexible minds, rigid rules: our model shows this contrast arises from a simple informational constraint—people often cannot tell whether noncooperation reflects genuine hardship or convenient excuses. This ambiguity gives opportunists room to exploit flexible rules. Rigid rules close that loophole, making cooperation more robust to strategic defection, even at the cost of sometimes enforcing inefficient behavior. This efficiency–robustness trade-off explains when societies tighten norms and why consequentialist moral reasoning can nonetheless give rise to judgments that seem deontological.

Materials and Methods

Model description.

Players. We consider a repeated game with $n \gg 1$ long-run signalers and an infinite pool of short-run choosers. Before play begins, signalers are assigned a discount factor δ according to a continuous distribution with full support on $\Delta = (0, 1)$. They privately observe their discount factor; the distribution is common knowledge.

Interactions. The game proceeds in two stages. Initially (round 0), signalers decide whether to contribute to a public good. Each contribution generates collective benefit β, divided equally among the $n - 1$ other signalers. The cost of contribution depends on circumstances: γ_H under hardship (probability $p \in (0, 1)$) and γ under normal conditions (probability $1 - p$), where $0 < \gamma < \beta < \gamma_H$. Signalers privately observe their cost before deciding.

Subsequently (rounds $t \geq 1$), each signaler plays a repeated trust game with a sequence of choosers. In every round, the signaler is paired with a new chooser, who decides whether to trust them, incurring cost c to provide benefit b (where $0 < c < b$). If trusted, the signaler chooses whether to reciprocate, incurring cost c to return benefit b.

Reputations. Signalers acquire a public goods game reputation ω_{pgg} and a trust game reputation ω_{tg}. Their full reputation is $\omega = (\omega_{\text{pgg}}, \omega_{\text{tg}}) \in \Omega$.

There are three possible values for ω_{pgg}. Signalers who contribute become contributors, those who free-ride under hardship become justified free-riders, and those who free-ride under normal conditions become justified free-riders with probability $\varepsilon \in (0, 1)$ and unjustified free-riders with probability $1 - \varepsilon$.

There are three possible values for ω_{tg}. Signalers who have not yet been trusted have unobserved trust game behavior; those who last reciprocated are reciprocators; and those who last cheated are cheaters.

For simplicity of exposition, when a signaler's trust game behavior is unobserved ($\omega_{\text{tg}} = $ unobserved), we refer to them by their public goods game reputation alone—for example, contributors are signalers with reputation (contributor, unobserved). Once a signaler has been trusted, we refer to them by their trust game reputation—for example, reciprocators are signalers with any reputation $(\cdot, \text{reciprocator})$.

Strategies and beliefs. A signaler public goods game strategy, represented by a function $\sigma_{\text{pgg}} : \Delta \times \{\text{normal, hardship}\} \rightarrow \{\text{contribute, free-ride}\}$, specifies whether to contribute depending on discount factor and circumstances. A signaler trust game strategy, represented by a function $\sigma_{\text{tg}} : \Delta \times \Omega \rightarrow \{\text{reciprocate, cheat}\}$, specifies whether to reciprocate trust depending on discount factor and reputation.

Choosers hold shared beliefs $\phi(\cdot \mid \omega)$ over Δ for each reputation $\omega \in \Omega$, encoding how they assign likelihood to different signaler discount factors, conditional on observed reputation. A chooser strategy, represented by a function $\sigma_{\text{ch}} : \Omega \rightarrow \{\text{trust, distrust}\}$, specifies whether to trust a signaler based on their reputation.

A strategy profile is $\sigma = (\sigma_{\text{pgg}}, \sigma_{\text{tg}}, \sigma_{\text{ch}})$.

Equilibrium analysis.

Equilibrium concept. We study the model's Perfect Bayesian Equilibria (PBEs) in pure strategies. A PBE is a pair (σ, ϕ) where players have no profitable deviations at any history (on or off the equilibrium path), and beliefs are updated according to Bayes' rule whenever possible ([75]).

Cooperative equilibria. We focus on cooperative PBEs, in which choosers' trust decisions incentivize cooperation at every history. Choosers must trust contributors and distrust unjustified free-riders (incentivizing contributions to the public good), and trust reciprocators and distrust cheaters (incentivizing reciprocation in every subgame). We obtain two cooperative PBEs: the flexible and the rigid norm.

In the Supplementary Information, we characterize the model's other PBEs. These fall into two categories: PBEs that are outcome-equivalent to the flexible or rigid norm, differing only in off-path play; and PBEs in which signalers never contribute to the public good.

Numerical solution.

Evaluation of domains and performance. We numerically evaluate the domain conditions for the flexible ([F2]) and rigid ([R3-R4]) norms on a parameter grid. These conditions are stated in terms of probabilities over discount factors; to evaluate them, we fix the distribution of discount factors to a normal with mean μ and standard deviation $\sigma = 0.3$, truncated to $(0, 1)$ (so μ refers to the pre-truncation mean). We vary impatience $1 - \mu$ and plausible deniability ε, holding other parameters fixed as in the figure captions. Figure 2 displays the domains of sustainable cooperation by identifying parameter combinations for which these conditions hold; Figure 3 compares performance where both norms are sustainable.

1. F Cushman, L Young, M Hauser, The role of conscious reasoning and intuition in moral judgment: testing three principles of harm. *Psychol. Sci.* **17**, 1082–1089 (2006).
2. JD Greene, RB Sommerville, LE Nystrom, JM Darley, JD Cohen, An fMRI investigation of emotional engagement in moral judgment. *Science* **293**, 2105–2108 (2001).
3. M Kleiman-Weiner, T Gerstenberg, S Levine, JB Tenenbaum, Inference of intention and permissibility in moral decision making in *Proceedings of the 37th Annual Conference of the Cognitive Science Society.* (Cognitive Science Society, Austin, TX), pp. 1123–1128 (2015).
4. M Kleiman-Weiner, R Saxe, JB Tenenbaum, Learning a commonsense moral theory. *Cognition* **167**, 107–123 (2017).
5. A Le Pargneux, F Cushman, Moral judgment is sensitive to bargaining power. *J. Exp. Psychol. Gen.* **154**, 263–278 (2025).
6. A Le Pargneux, N Chater, H Zeitoun, Contractualist tendencies and reasoning in moral judgment and decision making. *Cognition* **249**, 105838 (2024).
7. S Levine, M Kleiman-Weiner, L Schulz, J Tenenbaum, F Cushman, The logic of universalization guides moral judgment. *Proc. Natl. Acad. Sci. United States Am.* **117**, 26158–26169 (2020).
8. L Young, R Saxe, When ignorance is no excuse: Different roles for intent across moral domains. *Cognition* **120**, 202–214 (2011).
9. S Levine, M Kleiman-Weiner, N Chater, F Cushman, JB Tenenbaum, When rules are over-ruled: Virtual bargaining as a contractualist method of moral judgment. *Cognition* **250**, 105790 (2024).
10. J Kwon, T Zhi-Xuan, J Tenenbaum, S Levine, When it is not out of line to get out of line: The role of universalization and outcome-based reasoning in rule-breaking judgments (2023).
11. A Le Pargneux, X Roberts-Gaal, L Woodley, JD Greene, FA Cushman, Hard bargains and even splits: Fairness judgments track bargaining power across diverse cultures. *OSF Prepr.* (2025).
12. E Awad, et al., When is it acceptable to break the rules? Knowledge representation of moral judgements based on empirical data. *Auton. Agents Multi-Agent Syst.* **38**, 35 (2024).
13. M Jambon, JG Smetana, Moral complexity in middle childhood: children's evaluations of necessary harm. *Dev. Psychol.* **50**, 22–33 (2014).
14. J Marshall, K Wynn, P Bloom, Do children and adults take social relationship into account when evaluating people's actions? *Child Dev.* **91**, e1082–e1100 (2020) _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/cdev.13390.
15. J Marshall, et al., When not helping is nice: Children's changing evaluations of helping during COVID-19. *Dev. Psychol.* **59**, 953–962 (2023).
16. C Boehm, *Moral Origins: The Evolution of Virtue, Altruism, and Shame.* (Soft Skull Press, New York), (2012).
17. L Glowacky, S Lew-Levy, How small-scale societies achieve large-scale cooperation. *Curr. Opin. Psychol.* **44**, 44–48 (2022).
18. MJ Gelfand, JR Harrington, JC Jackson, The strength of social norms across human groups. *Perspectives on Psychol. Sci.* **12**, 800–809 (2017).
19. L Kohlberg, *Essays on Moral Development, Volume I: The Philosophy of Moral Development.* (Harper & Row, San Francisco), (1981).
20. JR Rest, Moral judgment research and the cognitive-developmental approach to moral education. *The Pers. Guid. J.* **58**, 602–605 (1980).
21. JR Snarey, J Reimer, L Kohlberg, Development of social-moral reasoning among kibbutz adolescents: A longitudinal cross-cultural study. *Dev. Psychol.* **21**, 3–17 (1985).
22. Southwark London Borough Council v Williams (1971).
23. J Altman, *Hunter-Gatherers Today: An Aboriginal Economy in North Australia*, Social Anthropology Series. (Australian Institute of Aboriginal Studies, Canberra) No. 7, (1987).
24. LJ Pospisil, *Kapauku Papuans and Their Law*, Yale University Publications in Anthropology. (Department of Anthropology, Yale University, New Haven) No. 54, (1958).
25. MS Alvard, DA Nolin, Rousseau's whale hunt? coordination among big-game hunters. *Curr. Anthropol.* **43**, 533–559 (2002).
26. D Damas, Central Eskimo systems of food sharing. *Ethnology* **11**, 220–240 (1972).
27. G Wenzel, Inuit subsistence and hunter support in nunavut in *Nunavut: Inuit Regain Control of Their Lands and Lives*, eds. J Dahl, J Hicks, P Jull. (International Work Group for Indigenous Affairs, Copenhagen), pp. 180–194 (2000).
28. J Dana, RA Weber, JX Kuang, Exploiting moral wiggle room: Experiments demonstrating an illusory preference for fairness. *Econ. Theory* **33**, 67–80 (2007).
29. Z Grossman, JJ Van Der Weele, Self-image and willful ignorance in social decisions. *J. Eur. Econ. Assoc.* **15**, 173–217 (2017).
30. L Aarøe, MB Petersen, Crowding out culture: Scandinavians and americans agree on social welfare in the face of deservingness cues. *J. Polit.* **76**, 684–697 (2014).
31. AP Fiske, The four elementary forms of sociality: Framework for a unified theory of social relations. *Psychol. Rev.* **99**, 689–723 (1992).
32. X Cheng, et al., The conceptual structure of human relationships across modern and historical cultures. *Nat. Hum. Behav.* pp. 1–14 (2025).
33. MJ Gelfand, et al., Differences between tight and loose cultures: A 33-nation study. *Science* **332**, 1100–1104 (2011).
34. JR Harrington, MJ Gelfand, Tightness–looseness across the 50 united states. *Proc. Natl. Acad. Sci.* **111**, 7990–7995 (2014).
35. JC Jackson, MJ Gelfand, CR Ember, A global analysis of cultural tightness and its relationship with ecological threat, social complexity, and social structure. *Proc. Royal Soc. B: Biol. Sci.* **287**, 20191227 (2020).
36. P Roos, M Gelfand, D Nau, J Lun, Societal threat and cultural variation in the strength of social norms: An evolutionary basis. *Organ. Behav. Hum. Decis. Process.* **129**, 14–23 (2015).
37. H Mell, N Baumard, JB André, Time is money: Waiting costs explain why selection favors steeper time discounting in deprived environments. *Evol. Hum. Behav.* **42**, 379–387 (2021).
38. D Nettle, R Saxe, 'if men were angels, no government would be necessary': The intuitive theory of social motivation and preference for authoritarian leaders. *Collabra: Psychol.* **7**, 28105 (2021).
39. J Duckitt, K Fisher, The impact of social threat on worldview and ideological attitudes. *Polit. Psychol.* **24**, 199–222 (2003).
40. L Laustsen, MB Petersen, Does a competent leader make a good friend? Conflict, ideology and the psychologies of friendship and followership. *Evol. Hum. Behav.* **36**, 286–293 (2015).
41. L Safra, et al., Childhood harshness predicts long-lasting leader preferences. *Evol. Hum. Behav.* **38**, 645–651 (2017).
42. KEG Williams, AM Votruba, SL Neuberg, MJ Saks, Capital and punishment: Resource scarcity increases endorsement of the death penalty. *Evol. Hum. Behav.* **40**, 65–73 (2019).
43. PE Tetlock, Thinking the unthinkable: Sacred values and taboo cognitions. *Trends Cogn. Sci.* **7**, 320–324 (2003).
44. JD Greene, JM Paxton, Patterns of neural activity associated with honest and dishonest moral decisions. *Proc. Natl. Acad. Sci.* **106**, 12506–12511 (2009).
45. J Haidt, SH Koller, MG Dias, Affect, culture, and morality, or is it wrong to eat your dog? *J. Pers. Soc. Psychol.* **65**, 613–628 (1993).
46. J Baron, I Ritov, Protected values and omission bias as deontological judgments in *Psychology of Learning and Motivation.* (Academic Press) Vol. 50, pp. 133–167 (2009).
47. JB André, L Fitouchi, S Debove, N Baumard, An evolutionary contractualist theory of morality (2022) Preprint.

48. JD Greene, *Moral tribes: Emotion, reason, and the gap between us and them.* (Penguin), (2014).

49. S Levine, N Chater, JB Tenenbaum, FA Cushman, Resource-rational contractualism: A triple theory of moral cognition. *Behav. Brain Sci.* (2024) Online ahead of print (received Oct 28, 2024).

50. E Yoeli, NA Dalkiran, B Burum, MA Nowak, M Hoffman, The emergence of categorical norms. *Preprint* (2022) Preprint available on ResearchSquare.

51. P DeScioli, R Kurzban, A solution to the mysteries of morality. *Psychol. Bull.* **139**, 477–496 (2013).

52. M Hoffman, E Yoeli, MA Nowak, Cooperate without looking: Why we care what people think and not just what they do. *Proc. Natl. Acad. Sci.* **112**, 1727–1732 (2015).

53. M Singh, M Hoffman, Commitment and impersonation: A reputation-based theory of principled behavior (2021) Preprint.

54. MA Nowak, K Sigmund, Evolution of indirect reciprocity by image scoring. *Nature* **393**, 573–577 (1998).

55. K Panchanathan, R Boyd, A tale of two defectors: the importance of standing for evolution of indirect reciprocity. *J. Theor. Biol.* **224**, 115–126 (2003).

56. O Leimar, P Hammerstein, Evolution of cooperation through indirect reciprocity. *Proc. Royal Soc. London. Ser. B: Biol. Sci.* **268**, 745–753 (2001).

57. H Ohtsuki, Y Iwasa, The leading eight: social norms that can maintain cooperation by indirect reciprocity. *J. Theor. Biol.* **239**, 435–444 (2006).

58. H Ohtsuki, Y Iwasa, How should we define goodness?—reputation dynamics in indirect reciprocity. *J. Theor. Biol.* **231**, 107–120 (2004).

59. Y Murase, C Hilbe, Computational evolution of social norms in well-mixed and group-structured populations. *Proc. Natl. Acad. Sci. United States Am.* **121**, e2406885121 (2024).

60. C Hilbe, L Schmid, J Tkadlec, MA Nowak, Indirect reciprocity with private, noisy, and incomplete information. *Proc. Natl. Acad. Sci.* **115**, 12241–12246 (2018).

61. FP Santos, FC Santos, JM Pacheco, Social norm complexity and past reputations in the evolution of cooperation. *Nature* **555**, 242–245 (2018).

62. J Lie-Panis, JB André, Cooperation as a signal of time preferences. *Proc. Royal Soc. B: Biol. Sci.* **289**, 20212266 (2022).

63. O Leimar, Reciprocity and communication of partner quality. *Proc. Royal Soc. London. Ser. B: Biol. Sci.* **264**, 1209–1215 (1997).

64. JB André, The evolution of reciprocity: Social types or social incentives? *The Am. Nat.* **175**, 197–210 (2010).

65. K Panchanathan, R Boyd, Indirect reciprocity can stabilize cooperation without the second-order free rider problem. *Nature* **432**, 499–502 (2004).

66. E Yoeli, M Hoffman, *Hidden Games: The Surprising Power of Game Theory to Explain Irrational Human Behavior.* (Basic Books, New York), (2022).

67. TA Kessinger, CE Tarnita, JB Plotkin, Evolution of norms for judging social behavior. *Proc. Natl. Acad. Sci.* **120**, e2219480120 (2023).

68. JM Pacheco, FC Santos, FACC Chalub, Stern-judging: A simple, successful norm which promotes cooperation under indirect reciprocity. *PLoS Comput. Biol.* **2**, e178 (2006).

69. M Kawakatsu, TA Kessinger, JB Plotkin, A mechanistic model of gossip, reputations, and cooperation. *Proc. Natl. Acad. Sci.* **121**, e2400689121 (2024).

70. D Nettle, C Heintz, Studying culture from an evolutionary psychological perspective: approaches, achievements, and prospects (PsyArXiv) (2025).

71. Y Murase, C Hilbe, Indirect reciprocity under opinion synchronization. *Proc. Natl. Acad. Sci.* **121**, e2418364121 (2024).

72. TA Kessinger, JB Plotkin, Institutions of public judgment established by social contract and taxation. *Proc. Natl. Acad. Sci.* **122**, e2506537122 (2025).

73. HP Young, The evolution of social norms. *Annu. Rev. Econ.* **7**, 359–387 (2015).

74. R Boyd, PJ Richerson, Punishment allows the evolution of cooperation (or anything else) in sizable groups. *Ethol. Sociobiol.* **13**, 171–195 (1992).

75. D Fudenberg, J Tirole, *Game Theory.* (MIT Press, Cambridge, MA), (1991).