

September 2026

“Cooperation thrives when moral judgment matches moral conduct”

Jorge Peña, Julian Garcia and Toby Handfield

Cooperation thrives when moral judgment matches moral conduct

Jorge Peña^{1,*}, Julian Garcia², and Toby Handfield³

¹Department of Social and Behavioral Sciences, Toulouse School of Economics, Toulouse, France

²Department of Data Science and AI, Faculty of Information Technology, Monash University, Melbourne, Australia

³School of Philosophical, Historical and International Studies, Monash University, Melbourne, Australia

*Corresponding author: jorge.pena@tse-fr.eu

A leading explanation for why humans cooperate with strangers is indirect reciprocity: when conduct is observed, individuals earn good reputations by cooperating with the deserving^{1,2}. Previous research has identified many ‘moral systems’ that sustain cooperation through indirect reciprocity – each pairing a strategy for how to act with a norm for how to judge others – yet what makes them succeed, and whether that success generalizes beyond the donation game, has remained unclear^{3–5}. Here we evaluate all 1,048,576 moral systems whose norms condition on both players’ reputations and actions, across every two-player social dilemma. We find that every highly stable, cooperative system pairs a ‘discriminating’ strategy – one that cooperates with the good and defects against the bad when in good standing – with a ‘consistent’ norm that deems good exactly those who act as the strategy prescribes. Such systems ‘practice what they preach’ and preach what they practice. In the limit of small errors, only one strategy paired with a consistent norm remains stable across all social dilemmas: the ‘contrite discriminator’, which cooperates to make amends when its standing is bad. Our results unify decades of research on indirect reciprocity, revealing moral consistency as the common ingredient of every stable, cooperative moral system.

Humans cooperate on a scale unmatched among other animals, routinely helping people they may never encounter again^{6,7}. A central explanation is reciprocity. Under direct reciprocity, individuals help those who have helped them⁸; under indirect reciprocity, they help those who have helped others^{9–11}, so that cooperation is rewarded not by second parties but by third parties who have observed or heard about one’s conduct¹. A reputation aggregates an individual’s past conduct into a standing that others can act on without any direct experience of them⁴. What turns such standings into cooperation is the ‘moral system’ adopted by society at large: the rules of

behavior and approval about “what is right and wrong”¹, typically formalized as a strategy paired with a social norm.

Prior work has uncovered an array of moral systems capable of sustaining cooperation through indirect reciprocity, the best known being the ‘leading eight’ first identified by Ohtsuki and Iwasa³ and studied extensively since^{12–16}. Yet this body of work has three important limitations. First, most of what we know about successful moral systems comes from a single, highly simplified interaction: the donation game, in which one player decides whether to pay a personal cost to confer a larger benefit on the other⁴. This is also the setting for almost all recent theoretical work, including studies of higher-order norms¹⁷, the evolution of norms^{18,19}, and indirect reciprocity under private reputations^{20–23}. Real-world human cooperation, by contrast, unfolds across a far wider range of situations. These certainly include prisoner’s dilemmas – of which the donation game is a special case – where cooperation is dominated. But they also include snowdrift (or chicken) games, where the best response to defection is to cooperate, and stag hunts (or assurance games), where each player is best off matching the other’s action^{24–26}. Second, previous analyses have each searched only a small subset of the possible social norms, ignoring some of the information available to observers in a single interaction^{3,5}. Third, and most fundamentally, we lack a unifying account of why some moral systems succeed. Existing characterizations are disparate and tied to particular settings, leaving no parsimonious explanation for what separates the moral systems that sustain cooperation from those that fail. Even the study that covered the largest range of games concluded that its successful norms “have different principles in different dilemmas for suppressing cheaters”⁵.

Here we address all three limitations at once. As in most models of indirect reciprocity^{3,4}, we take actions and reputations to be binary: an individual either cooperates or defects, and is judged either good or bad. We study the complete space of social norms that condition on the two players’ reputations and actions. Combined with the $2^4 = 16$ possible strategies, this yields $2^{20} = 1,048,576$ moral systems, which we evaluate across the entire set of symmetric two-player cooperative dilemmas where mutual cooperation is socially optimal²⁴. Our analysis reveals a single organizing principle behind every successful moral system: whatever the dilemma, those systems that best sustain cooperation pair a ‘discriminating’ strategy, which conditions cooperation on the partner’s reputation, with a ‘consistent’ social norm, which awards a good reputation to exactly those who act as the strategy dictates. Among these consistent systems, those built on a ‘contrite’ strategy – one that cooperates to make amends when its own standing is bad – stand out as the only moral systems that remain both stable and cooperative across every

class of dilemma at once.

Results

Mapping the full space of moral systems

Modeling indirect reciprocity requires three ingredients ([Methods](#)): a payoff matrix, which defines the rewards and costs of cooperating and defecting (Fig. 1a); a strategy (or action rule), which specifies when to cooperate and when to defect (Fig. 1b); and a social norm (or assessment rule), which tells observers when to assign a good or a bad reputation (Fig. 1c). Strategies and social norms can be classified by the amount and nature of the information they take as input. For instance, early models considered a moral system where the strategy ('discriminator' or DISC; Fig. 1b) prescribes cooperation if the co-player's reputation is good and defection if the co-player's reputation is bad, and the social norm ('Scoring'; Fig. 1c) assigns a good reputation if the player cooperated and a bad reputation if the player defected². Both the strategy and the social norm of this moral system require one bit of information, and are thus said to be of first order. Although this moral system can sustain cooperation against unconditional defectors, it is not robust to invasion by other strategies²⁷.

Ohtsuki and Iwasa³ were the first to systematically study all moral systems obtained by pairing each of the $2^4 = 16$ possible strategies of up to second order (based on two bits of information, namely the reputations of the focal player and the co-player) with each of the 2^8 possible social norms of up to third order (based on three bits of information, namely the action of the player being assessed and the reputations of player and co-player). Out of these $2^{12} = 4,096$ moral systems, only eight – known as the 'leading eight' – were found both to sustain high levels of cooperation and to be evolutionarily stable when social interactions were modeled as a donation game. Two examples of the leading eight are L6 and L1 (labels adopted from²⁰). L6 employs DISC together with a second-order social norm ('Stern Judging'; Fig. 1c) that calls an individual good only if they cooperated with a good partner or defected against a bad one. In contrast, L1 uses a second-order strategy that takes the individual's own reputation into account, prescribing cooperation when the co-player is good or when one is oneself bad. This prescription to cooperate when in bad standing can be read as a form of contrition that attempts to repair one's damaged reputation¹²; we therefore call this strategy 'contrite discriminator' (CDISC; Fig. 1b). L1 pairs CDISC with the third-order social norm 'Standing' (Fig. 1c), which also conditions on the reputation of the individual being judged: good individuals keep their good reputation unless

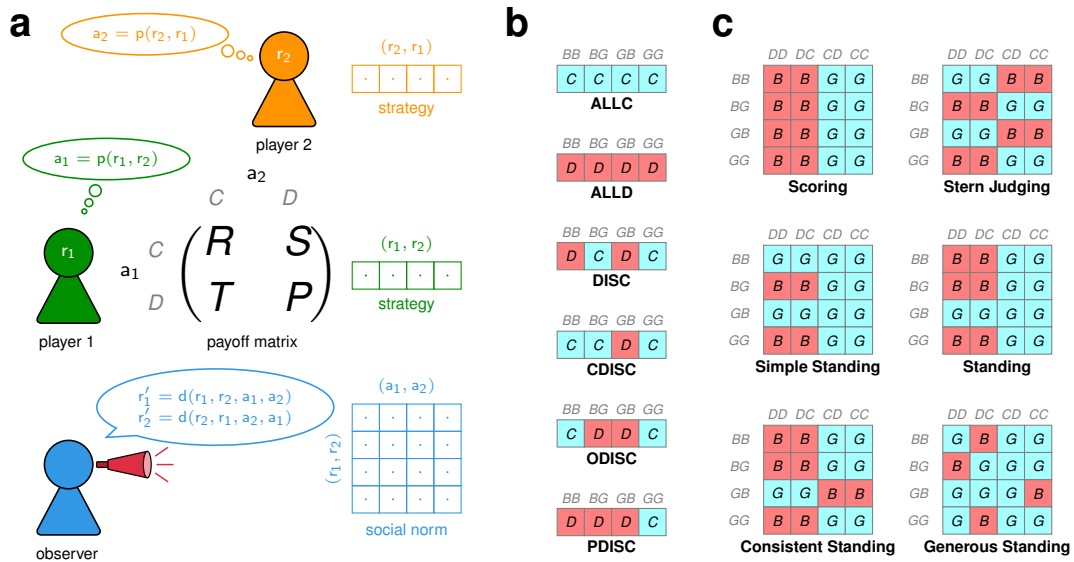


Figure 1: Modeling indirect reciprocity. **a**, Player 1 (green pawn, reputation $r_1 \in \{G, B\}$, good or bad) and player 2 (orange pawn, reputation r_2) play a game in which each can cooperate (C) or defect (D). Player 1 chooses her action $a_1 = p(r_1, r_2)$ according to her strategy p (green table); player 2 likewise chooses $a_2 = p(r_2, r_1)$ (orange table). The players act simultaneously and obtain the payoffs shown in the matrix (each entry gives the row player's payoff), whose ordering defines the class of cooperative dilemma: prisoner's dilemma, snowdrift game, or stag hunt (Methods; Extended Data Fig. 2). An observer (blue pawn) witnesses the interaction and assigns new reputations – $r'_1 = d(r_1, r_2, a_1, a_2)$ and $r'_2 = d(r_2, r_1, a_2, a_1)$ – according to the social norm d (blue table). Assessment is public: reputations are shared by the whole population (Methods). **b**, A strategy is a table whose entries give player 1's action for each combination of the players' reputations (r_1, r_2); shown are the unconditional cooperator (ALLC) and defector (ALLD), and the four discriminating strategies DISC, CDISC, ODISC, and PDISC. **c**, A social norm is a table whose entries give player 1's new reputation for each combination of reputations (r_1, r_2) and actions (a_1, a_2); shown are Scoring, Stern Judging, Simple Standing, Standing, Consistent Standing, and Generous Standing.

they defect against a good partner, while bad individuals turn good if and only if they cooperate, whatever the partner's reputation.

Within indirect reciprocity, social dilemmas beyond the donation game have rarely been studied. A notable exception is the study by Nakamura and Ohtsuki⁵, who analyzed indirect reciprocity across prisoner's dilemmas, snowdrift games, and stag hunts, and identified which social norms stabilize cooperation in each game. Building on their work, we ask whether a single moral system can sustain cooperation across the entire range of cooperative dilemmas.

Prior analyses differ in how much of the information available in a single interaction their social norms use (Extended Data Fig. 1). The norms studied by Ohtsuki and Iwasa^{3,12} condition on the focal player's action and on both players' reputations, but never on the co-player's action; the norms of Nakamura and Ohtsuki⁵ condition on both players' actions and the co-player's

reputation, but never on the focal player’s own reputation. Norms that use all four pieces of information available in a single interaction – the two players’ reputations and actions – have so far been left unexplored⁵. We study the complete fourth-order space of $2^{16} = 65,536$ social norms, which, paired with the 16 strategies of up to second order, yields the $2^{20} = 1,048,576$ moral systems we evaluate across the entire set of symmetric 2×2 cooperative dilemmas where mutual cooperation is socially optimal (Extended Data Fig. 2). Throughout, we adopt a public-assessment model of reputation: a single observer witnesses each interaction and judges both players, with the reputations they assign shared by the whole community, so that everyone comes to hold the same view of any given individual^{3,28}.

Highly stable systems are consistent

We characterize each moral system by two quantities: its cooperation index and its stability index (see Methods). The cooperation index (κ) is the probability that a randomly chosen resident cooperates at equilibrium. Because this probability does not depend on the payoffs, it is the same for every game we study. The stability index for a given dilemma class is the fraction of payoff space over which the resident strategy resists invasion by any alternative strategy, holding fixed the social norm. We then take the simple average across the three dilemma classes to obtain the average stability index (σ). We report results for low error rates in the execution of cooperation ($\epsilon = 0.02$) and in the assignment of reputations ($\chi = 0.02$), but our conclusions are robust to other small error rates and to alternative constructions of the stability index (SI Figs. 5, 26, 27).

Computing the cooperation index and the average stability index for all 2^{20} moral systems – where systems that differ only by swapping the two reputation labels, and thus prescribe identical behavior, are evaluated only once (SI Section 4.1) – reveals our main result (Fig. 2): a single cluster of $2^{10} = 1,024$ moral systems stands out as being both highly stable and highly cooperative, separated from the rest by a clear gap in stability. The moral systems in this cluster share two properties. First, they each use a ‘discriminating’ strategy: players in good standing cooperate with good co-players and defect against bad ones (the four discriminating strategies – DISC, CDISC, ODISC, and PDISC – are shown in Fig. 1b; Methods). Second, they pair such a discriminating strategy with a ‘consistent’ social norm: one that, conditional on the co-player playing the same strategy, assigns a good reputation to all and only those who behave as the system’s own strategy prescribes (Fig. 3a; Methods). Because consistency says nothing about cases where the co-player deviates from the system’s strategy, it leaves eight bits unconstrained, so each strategy admits $2^8 = 256$ consistent norms (for CDISC, these include, for instance, Consistent Standing and

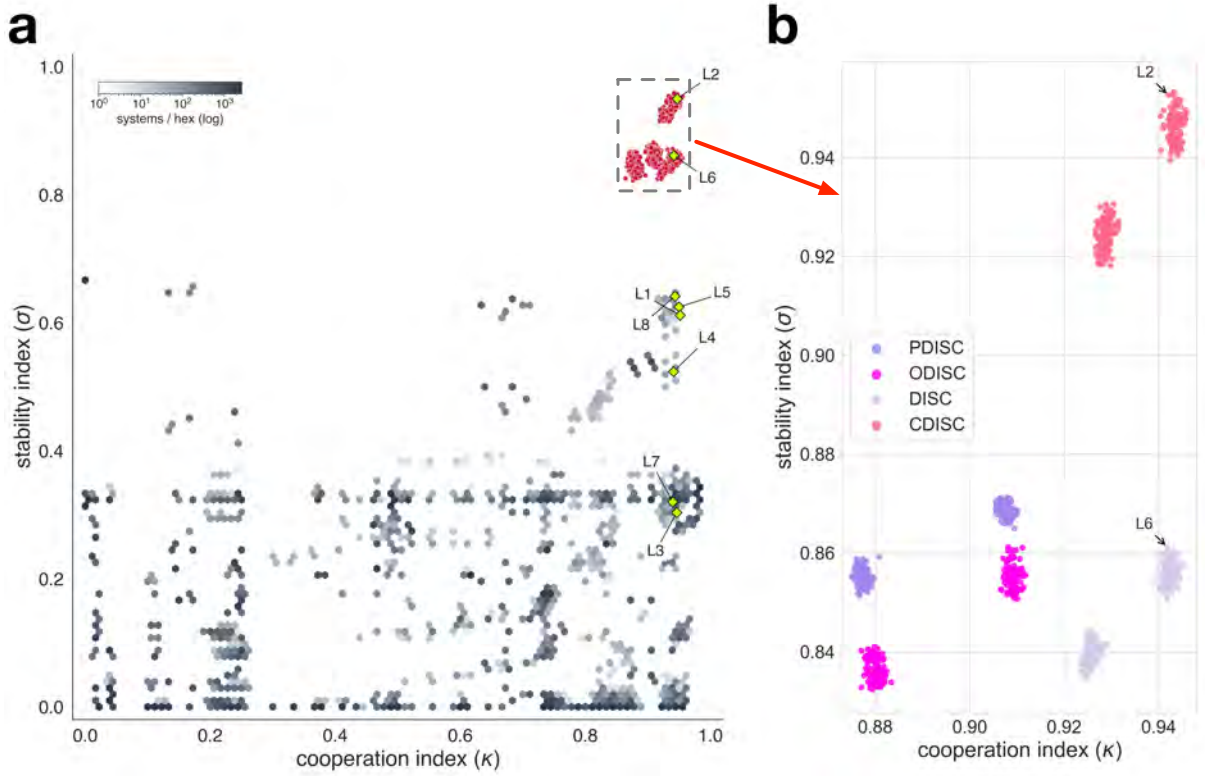


Figure 2: Stability and cooperation indices for all moral systems. **a**, Each of the 524,800 moral systems that are distinct up to the relabeling of reputations (one representative per mirror class; SI Section 4.1) is positioned by its resident cooperation index (κ , horizontal axis) and average stability index (σ , vertical axis); here $\chi = \epsilon = 0.02$. The background hexbin layer shows the density of the 523,776 representatives that are not consistent discriminating systems, on a logarithmic color scale (inset). The 1,024 consistent discriminating systems (red), shown as a jittered scatter plot overlaid on the density map, cluster tightly in the upper-right corner, separated from the bulk of moral systems by a clear gap in stability. Of the leading-eight moral systems (L1–L8, green diamonds), only L2 and L6 fall within this cluster. **b**, Zoomed-in scatter plot of the consistent discriminating systems (jittered for clarity), colored by strategy; the 256 systems using CDISC (the ‘consistent contrite discriminating’ systems) achieve the highest stability. See SI Figs. 5–22 for results at alternative error values.

Generous Standing; Fig. 1c). The four discriminating strategies thus give rise to the 1,024 systems of the cluster, which we call ‘consistent discriminating’. Intuitively, consistency can be likened to moral integrity and an absence of hypocrisy: the standard to which one holds others is exactly the standard by which one acts. Interestingly, all leading-eight moral systems use discriminating strategies (DISC and CDISC), but only two are consistent: L2 and L6 (Fig. 3b).

The success of consistent discriminating systems follows from the basic function of a moral system in indirect reciprocity: namely, to sort the population so that reciprocators can direct cooperation toward individuals with the same strategy, and defect against all others. Indeed, a consistent social norm is the one best suited to this task, as it assigns reputations so that those

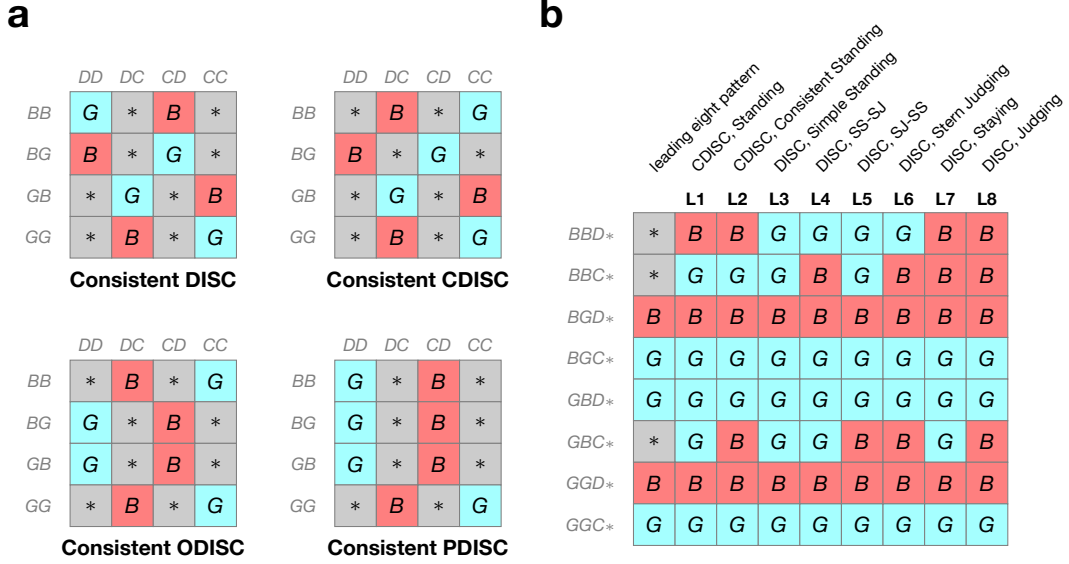


Figure 3: **Social norms of the consistent discriminating systems and of the leading eight.**

a, The social norm classes consistent with each of the four discriminating strategies (Fig. 1b); because consistency constrains a norm only when the co-player behaves in accordance with the resident strategy, eight of the sixteen bits are left unconstrained (*), so each strategy admits $2^8 = 256$ consistent norms. **b**, The social norms of the leading-eight moral systems. Rows are labeled by the reputations and action ($r_1 r_2 a_1$) of the judged player, with * standing for the co-player's action a_2 , on which these third-order norms do not depend; the shared pattern column shows the entries common to all eight, with * marking the three rows on which they differ. Only two of the leading eight feature a consistent social norm: L2 (CDISC, Consistent Standing) and L6 (DISC, Stern Judging).

who follow the resident strategy are labeled ‘good’, and those who deviate are labeled ‘bad’. An inconsistent norm, by contrast, will assign good reputations to some mutant strategies at least as often as to residents, blurring the very distinction on which discrimination depends, thus opening the door to invasion. We can make this intuition precise in the absence of execution errors: a moral system is consistent if and only if, at equilibrium, residents are judged good strictly more often than any mutant that deviates from the resident strategy (SI Propositions 8, 9).

Contribution makes stability universal

Within the cluster of consistent discriminating systems, a further subdivision emerges according to whether the strategy is contrite (Fig. 2b). A ‘contrite’ strategy prescribes cooperation whenever one’s own standing is bad (Methods); among the discriminating strategies, CDISC is the only one with this property. The 256 consistent systems associated with this contrite discriminator achieve the highest stability of any moral system for all but the largest assignment error rate we examine (Fig. 2b; SI Fig. 6), remaining highly stable in every class of dilemma (SI Section 4.4). We

further examine this finding analytically in a tractable regime: no execution errors ($\epsilon = 0$) and the limit of vanishing assignment errors ($\chi \rightarrow 0$). Of the 2^{20} moral systems we investigate, 114,688 are ‘righteous’: those in which residents both cooperate at near-maximal rates and almost all attain good standing (Methods; SI Definition 11, Proposition 5). In structural terms, a righteous strategy is reputation-sensitive and prescribes cooperation with a good partner when in good standing. A righteous norm, assessing a resident whose partner also follows the strategy, judges that resident good in at least three of the four combinations of the two players’ standings – always when both are good, and in at least two of the remaining three (SI Eq. (80)). The 1,024 consistent discriminating systems and all leading-eight systems are righteous (SI Propositions 7, 10). We ask which righteous systems are ‘universally stable’, that is, resistant to invasion across every cooperative dilemma we consider. We show that a righteous moral system is universally stable if and only if its strategy is CDISC and its social norm is consistent (SI Theorem 1).

We identify the universally stable systems by discarding, in turn, the righteous systems invadable by three important classes of strategies (SI Section 3.5; SI Fig. 4). First, resistance to unconditional defectors (ALLD) requires two conditions. The strategy must prescribe defection against a bad partner when in good standing; otherwise residents behave indistinguishably from unconditional cooperators and ALLD invades directly in prisoner’s dilemmas and snowdrift games. All discriminating strategies meet this condition. The social norm must also judge (mutual) defection against a good partner, when in bad standing, as bad (SI Propositions 12, 11). Among the consistent discriminating systems, those with DISC and CDISC as strategies satisfy this condition, whereas those with ODISC and PDISC do not (SI Proposition 14). All leading-eight systems satisfy this condition as well (SI Proposition 13). Overall, 12,800 righteous systems resist unconditional defectors across every dilemma.

Second, resistance to unconditional cooperators (ALLC) is the converse problem. Since a discriminator in good standing defects only against bad partners, and almost all residents hold good reputations, an ALLC mutant differs from a resident mainly by cooperating with the rare bad ones. The social norm must therefore treat this unwarranted cooperation as a moral failure (SI Propositions 16, 17, 18). The consistent systems built on DISC and CDISC satisfy this condition, but only four of the leading eight – L2, L5, L6, and L8 – do; L2 and L6 are among those consistent systems, while L5 and L8 are not. The other four leading-eight systems (L1, L3, L4, and L7) praise cooperation with a bad partner and are invaded by ALLC in stag hunts (SI Propositions 20, 19; Fig. 4b). Overall, only 4,608 righteous systems resist both unconditional defectors and unconditional cooperators across every dilemma.

Finally, resistance to mutant discriminating strategies singles out CDISC under a consistent norm – that is, the consistent contrite discriminating systems. The crucial asymmetry is that CDISC can invade DISC under a consistent norm, but not vice versa. The two strategies act differently only when both players are bad: DISC defects, whereas CDISC shows contrition and cooperates. A consistent DISC norm judges this contrition bad, so a CDISC mutant needs two rounds to regain good standing, collecting $2S$ where a DISC resident collects $P + R$ (Extended Data Fig. 3); whenever $2S > R + P$, as is possible in snowdrift games, CDISC invades. Among CDISC residents under a consistent norm, a DISC mutant defecting in the rare state where both players are bad earns T but is judged bad, and collects $T + S$ over the two rounds it takes to regain good standing; a resident in the same state cooperates, is immediately judged good, and collects $2R$ (Extended Data Fig. 4). Because mutual cooperation is efficient ($2R > T + S$), the mutant has no margin to invade.

CDISC under a consistent norm resists not only rival discriminators but also every alternative mutant strategy (SI Theorem 1). This is our central analytical result: with no execution errors and in the limit of vanishing assignment errors, a righteous moral system is evolutionarily stable across all cooperative dilemmas where mutual cooperation is efficient if and only if its strategy is discriminating and contrite, and its social norm is consistent with it – the 256 consistent contrite discriminating systems. Only one member of the leading eight belongs to this set: L2 (CDISC, Consistent Standing). In this sense, we can call L2 the ‘leading one’. L2 is also the simplest consistent contrite discriminating system, as its third-order social norm ignores the co-player’s action.

Uncontrite and inconsistent systems fail

Knowing which consistent discriminating systems are universally stable, we can also ask precisely what goes wrong for the rest (Fig. 4). DISC under a consistent norm is stable across all prisoner’s dilemmas and stag hunts, yet, as already mentioned, is invaded by CDISC in snowdrift games with $2S > R + P$ (yielding $\sigma = 8/9$). The remaining two discriminators, ODISC and PDISC, exhibit the complementary vulnerability: under consistent norms, they are stable across all snowdrift games and stag hunts, but invadable in prisoner’s dilemmas with $2R < T + P$ (again $\sigma = 8/9$).

For the leading eight, the analytical results reveal which specific mutant strategies can invade and in which game classes (SI Section 3.7). We find that all leading-eight systems are stable across the entire class of prisoner’s dilemmas, and not merely the donation game assumed by the original analysis³. However, when we expand the analysis to the full space of cooperative dilemmas, L2

		cooperation index, κ	stability index, σ PD	stability index, σ SG	stability index, σ SH	average stability index, σ
a	Consistent DISC	1	1	2/3	1	8/9
	Consistent CDISC	1	1	1	1	1
	Consistent ODISC	1	2/3	1	1	8/9
	Consistent PDISC	1	2/3	1	1	8/9
b	L1 (CDISC, Standing)	1	1	1	0	2/3
	L2 (CDISC, Consistent Standing)	1	1	1	1	1
	L3 (DISC, Simple Standing)	1	1	0	0	1/3
	L4 (DISC, SS-SJ)	1	1	2/3	0	5/9
	L5 (DISC, SJ-SS)	1	1	0	1	2/3
	L6 (DISC, Stern Judging)	1	1	2/3	1	8/9
	L7 (DISC, Staying)	1	1	0	0	1/3
	L8 (DISC, Judging)	1	1	0	1	2/3

Figure 4: **Cooperation and stability of consistent discriminating systems and of the leading eight.** The cooperation index and the stability indices across the three classes of cooperative dilemma (prisoner’s dilemma, PD; snowdrift game, SG; and stag hunt, SH), with no execution errors ($\epsilon = 0$) and in the limit of vanishing assignment errors ($\chi \rightarrow 0$); cells are shaded by the value of the index, darker for lower values. **a**, The four classes of consistent discriminating systems (their norms are shown in Fig. 3a); in this limit, all 256 members of a class share the same indices. **b**, The leading-eight moral systems, each labeled by its strategy and social norm (their norms are shown in Fig. 3b).

and L6 are the only ones that fare particularly well, because they are the only leading-eight moral systems that feature a consistent social norm. The remaining members of the leading eight all fail to achieve high levels of stability in snowdrift games, stag hunts, or both (Fig. 4b).

Discussion

The moral systems that succeed under indirect reciprocity have been described in many ways – as those that distinguish ‘justified’ from ‘unjustified’ defection^{3,5}, or that rely on social norms that are ‘simple and unambiguous’ and ‘in which forgiveness and repent are welcome, while punishment is implacable’²⁹. We have shown that what these successful systems share is moral consistency: the social norm deems good exactly those who act as the strategy prescribes. For all two-player cooperative dilemmas with efficient mutual cooperation, and across a wide range of error rates, consistent discriminating systems are both highly stable and highly cooperative. The

underlying mechanism is easy to state: a discriminating system whose social norm is inconsistent will award good reputations to some mutant at least as often as to its own residents, misdirecting cooperation and becoming vulnerable to invasion. In the limit of small errors, consistent contrite discriminating systems are the only moral systems that are universally stable. Contrition – the readiness to cooperate when in bad standing – is what secures stability in the snowdrift games where a consistent but uncontrite discriminator can be invaded.

Why did earlier literature not identify consistency and contrition as key ingredients of a successful moral system? Part of the answer lies in the limited scope of earlier analyses, which considered a restricted set of invaders, a restricted range of social dilemmas, or a restricted space of social norms. Because each dilemma poses its own profile of invasion risks, an analysis restricted to a subset of social dilemmas subjects moral systems to a weaker stability test: an inconsistent system that resists invasion in one dilemma can still be invaded in another. Nakamura and Ohtsuki's analysis⁵ comes closest to ours: they surveyed the full range of dilemmas and allowed strategies to condition on the focal player's own reputation, the contrite discriminator among them. Yet none of their moral systems was both stable and cooperative across dilemmas. The limitation lay in the set of norms they explored. Because a contrite discriminator cooperates with a bad partner only when it is itself in bad standing, a consistent norm must reward that act in a bad player while condemning it in a good one, a distinction it can draw only by reading the focal player's own reputation. Nakamura and Ohtsuki's norms conditioned on both players' actions and the co-player's reputation but not the focal player's own reputation, and so could not draw this distinction.

Which moral systems appear successful depends on the range of dilemmas used to test them. Consequently, the catalogues of 'good' norms inherited from earlier work are, to some degree, artifacts of the kinds of social dilemma used to model social interactions. Our results thus call for a reappraisal of the best known of these, the leading eight^{3,12}. Once we move to the full range of cooperative dilemmas, this set proves at once too inclusive and too exclusive. It is too inclusive because only two of its members – L2 and L6 – are consistent; the remaining six can be invaded once we consider social interactions beyond the prisoner's dilemma. It is too exclusive because it admits only the DISC and CDISC strategies, leaving out ODISC and PDISC. However, consistent systems built on these excluded strategies are stable across all snowdrift games and stag hunts (Fig. 4a), and as a result achieve a higher average stability index than any inconsistent member of the leading eight – even though leading-eight systems remain more stable within the class of prisoner's dilemmas.

The consistent ODISC and PDISC systems deserve closer attention, because the behavior they prescribe resonates with patterns seen in human societies. In particular, PDISC encodes a social logic that evokes the phenomenon of honor cultures – societies in which retaliation for insult is strongly socially endorsed, and in which failure to answer a slight can itself attract contempt^{30–32}. Honor cultures are typically organized around a mixture of direct and indirect reciprocity: people retaliate against the particular individual who wronged them, often under public conditions in which status is at stake. By contrast, our model is a pure case of indirect reciprocity and does not permit targeting a specific offender. Even so, a consistent PDISC system can be read as a formal echo of honor culture. When its standing is compromised, a PDISC individual defects regardless of the partner’s reputation – and under its consistent norm this very defection is judged good, so that standing is restored by adversarial display rather than by apology. Viewed through this lens, honor cultures may not be pathologies of reciprocity but one among several consistent equilibria in which a moral system can settle and sustain cooperation.

More broadly, our results speak to a substantial body of empirical work on human indirect reciprocity. Controlled experiments show that people cooperate more with recipients of good reputation, and give more generously when their own conduct is observed and may be reported to future partners^{33–35}; likewise, reputation shapes cooperation across a wide range of natural social settings³⁶. A recurring question in this literature is which social norm people actually apply when judging others^{33,37–40}. Our theory sharpens that question. It predicts that a population should sustain a norm consistent with the strategy it plays, so that the standard people hold others to coincides with the behavior they display themselves. This alignment of judgment and conduct – and how it might break down when information becomes private or noisy – offers a target for empirical tests of when to expect reputation-based cooperation.

Our analysis assumes that individuals apply the same moral system whatever game they happen to be in. An alternative is to make behavior contingent on the game itself. Colnaghi et al.⁴¹ show that a costly ability to infer the structure of a social interaction and adapt to it can evolve and promote cooperation when the mix of games is sufficiently varied. Transposed to indirect reciprocity, this would require an entire society to detect which game it is playing and switch its moral system in concert. For instance, a population could adopt L3 in prisoner’s dilemmas, L1 in snowdrift games, and L5 in stag hunts (each stable within its own class, for an average stability index of one). Yet every moral system in such a repertoire is inconsistent. Our results point to a complementary possibility. Any consistent contrite discriminating system is stable across the entire range of cooperative dilemmas, and thus needs no ability to identify the game and incurs

no cost of switching. The whole work of sustaining cooperation can thus fall to a single moral system.

We have worked throughout with a public model of reputation, in which a single shared assessment of each individual circulates through the population. This is a substantive assumption, relaxed by a growing body of work asking what becomes of cooperation when observers form their own private and sometimes conflicting opinions^{20–23,28,42}. We adopt the public case not because private assessment is unimportant, but because it is the cleanest setting in which to study how social norms make reputation a reliable signal on which to condition behavior. Real populations surely draw on both shared and private information⁴³, and tracing how moral consistency fares as assessment grows more private is a natural and important extension of the present work.

We have shown that the moral systems that sustain cooperation across most social dilemmas are discriminating and consistent, and that those that prevail across every dilemma are contrite as well. Consistency is a close cousin of the golden rule, the principle – found in some form across many moral traditions – that we should treat others as we would have them treat us. But where the golden rule governs conduct, consistency governs judgment: it asks us to hold others to the standard we keep ourselves. For the contrite discriminator, consistency pairs a rule of conduct – cooperate with everyone, but punish a bad partner when in good standing – with a matching rule of judgment applied to others: that same punishment is deemed good in a partner, and those who cooperate while in bad standing regain their good name. Across the entire range of cooperative dilemmas, selection favors this alignment of judgment and conduct – moral systems that practice what they preach, and that preach what they practice.

Methods

Model

Cooperative dilemmas. We model social interactions as symmetric 2×2 games (Fig. 1a). The action of a focal player (a_1) and that of their co-player (a_2) can be either ‘cooperate’ (C) or ‘defect’ (D). We denote the payoffs for mutual cooperation, unilateral cooperation, unilateral defection, and mutual defection respectively by R , S , T , and P .

We focus on games that constitute social dilemmas of cooperation, or ‘cooperative dilemmas’, in the following sense²⁴. First, actions C and D qualify as cooperation and defection: (i) mutual cooperation is preferred over mutual defection ($R > P$), and (ii) each player prefers their co-player

to cooperate rather than defect, whatever their own action ($R > S$ and $T > P$). Second, the game is a genuine social dilemma: it has a (possibly mixed) evolutionarily stable strategy that does not coincide with the social optimum – the profile maximizing total payoff – which is the case if and only if $T > R$ or $P > S$. Hence, depending on the ranking of payoff values, there are three classes of cooperative dilemma: (i) prisoner’s dilemmas (PD) if $T > R > P > S$; (ii) snowdrift games (SG, also known as chicken games) if $T > R > S > P$; and (iii) stag hunts (SH, also known as assurance games) if $R > T > P > S$. The donation game used in most previous studies is a special case of the prisoner’s dilemma, with $R = b - c$, $S = -c$, $T = b$, $P = 0$, where the benefit of cooperation b and the cost of cooperation c satisfy $b > c > 0$. We further restrict attention to cooperative dilemmas where mutual cooperation is efficient, that is, where it is itself the social optimum; a sufficient condition for this, which we assume throughout, is $2R > T + S$ ²⁴.

Because our analyses are invariant under positive affine transformations of the payoffs – adding a common constant and multiplying by a positive one – we may normalize them by the min-max transformation

$$\Pi' = \frac{\Pi - \min \{R, S, T, P\}}{\max \{R, S, T, P\} - \min \{R, S, T, P\}}, \quad \text{for all } \Pi \in \{R, S, T, P\}, \quad (1)$$

which maps the largest payoff to one and the smallest to zero. This fixes two of the four payoffs, leaving each cooperative dilemma class $\Gamma \in \{\text{PD}, \text{SG}, \text{SH}\}$ with two free parameters; each class can then be represented as a polygon whose points are the individual dilemmas of that class (Extended Data Fig. 2), a representation we exploit when computing stability indices (see below). Our analysis uses the following normalized dilemmas, for which the efficiency condition $2R > T + S$ reduces to the constraint indicated: (i) $1 > R > P > 0$ with $R > 1/2$ (PD); (ii) $1 > R > S > 0$ with $2R > 1 + S$ (SG); and (iii) $1 > T > P > 0$ with no further constraint (SH, where it holds automatically). The normalized donation game is the special case of the normalized prisoner’s dilemma with $P = 1 - R$.

Strategies and execution errors. Strategies are used by individuals when deciding how to play – when to cooperate with, and when to defect against, a co-player. The action of an individual in each interaction depends on their own reputation (r_1) and on that of the co-player (r_2). With binary reputations (‘good’ = $G = 1$ or ‘bad’ = $B = 0$), the strategy used by each individual is a Boolean function $p : \{B, G\}^2 \rightarrow \{D, C\}$ prescribing an action (‘cooperate’ = $C = 1$ or ‘defect’ = $D = 0$) to each combination of these two reputations. We denote each strategy by a string $p = p_0 p_1 p_2 p_3$, where the bit $p_i \in \{D, C\}$ gives the action prescribed for one combination

of reputations, ordered r_1 then r_2 with B before G : p_0 corresponds to $r_1 = r_2 = B$, p_1 to $r_1 = B$ and $r_2 = G$, p_2 to $r_1 = G$ and $r_2 = B$, and p_3 to $r_1 = r_2 = G$. Such strategies are exactly the Boolean functions of two arguments, of which there are $2^{2^2} = 2^4 = 16$.

We model execution errors, which capture an individual's occasional inability to carry out an intended action. As in previous work^{3,5}, these errors affect only intended cooperation: an individual who intends to cooperate defects instead with probability $\epsilon \in [0, 1/2)$ (owing, for instance, to a lack of ability, time, or energy), whereas intended defection is always carried out.

Social norms, assignment errors, and observation model. Social norms are used by observers when deciding how to assign reputations to individuals after a social interaction. We model the social norm as a Boolean function $d : \{B, G\}^2 \times \{D, C\}^2 \rightarrow \{B, G\}$ prescribing a new reputation ('good' = $G = 1$ or 'bad' = $B = 0$) to each combination of the reputations and actions of the individual being judged and their co-player. We denote each social norm by a string $d = d_0 d_1 \dots d_{15}$, where the bit $d_i \in \{B, G\}$ gives the new reputation assigned to the judged individual for one combination of the current reputations r_1, r_2 and actions a_1, a_2 of that individual and their co-player, ordered r_1, r_2, a_1, a_2 with B before G and D before C . For example, d_0 corresponds to $r_1 = r_2 = a_1 = a_2 = 0$, d_5 to $r_1 = a_1 = 0$ and $r_2 = a_2 = 1$, and d_{15} to $r_1 = r_2 = a_1 = a_2 = 1$. Each social norm can equivalently be displayed as a table, as in Figs 1c and 3a, whose rows list the reputation pairs $(r_1, r_2) = BB, BG, GB, GG$ from top to bottom and whose columns list the action pairs $(a_1, a_2) = DD, DC, CD, CC$ from left to right. The bitstring $d_0 d_1 \dots d_{15}$ is simply this table read row by row: d_0 is the top-left entry (reputations BB , actions DD) and d_{15} the bottom-right one (reputations GG , actions CC). Such social norms are exactly the Boolean functions of four arguments, of which there are $2^{2^4} = 2^{16} = 65,536$, spanning different orders and classes (see [Extended Data Fig. 1](#)).

We also model assignment errors: with probability $\chi \in (0, 1/2)$, the observer assigns the reputation opposite to the one prescribed by the social norm. We adopt a 'public-assessment model'^{3,12,13,17}: once an individual's reputation is updated, it is transmitted to the whole population (for instance, through gossip or a central authority), so that everyone agrees on the reputation of others.

Moral systems. A moral system (p, d) pairs a strategy p with a social norm d ; a population implements it if, at all times, individuals behave according to p and observers judge according to d . Combining every strategy with every social norm yields $2^4 \times 2^{16} = 2^{20} = 1,048,576$ moral systems.

Classes of moral systems

Discriminating strategies and systems. Among the sixteen strategies, those that cooperate with a good co-player and defect against a bad one when the focal player is in good standing ($r_1 = G$) are central to our analysis. Formally, a strategy p is ‘discriminating’ if $p(G, B) = D$ and $p(G, G) = C$; as a string, $p = **DC$. These two fixed bits leave behavior in bad standing unspecified, yielding the four discriminating strategies (Fig. 1b), which differ only in what they prescribe in that case: the ‘(standard) discriminator’ (DISC, $p = DCDC$) discriminates by the co-player’s reputation just as it does when in good standing; the ‘contrite discriminator’ (CDISC, $p = CCDC$) cooperates regardless of the co-player’s reputation; the ‘obfuscated discriminator’ (ODISC, $p = CDDC$) reverses the discriminating logic; and the ‘petulant discriminator’ (PDISC, $p = DDDC$) defects regardless of the co-player’s reputation. A moral system (p, d) is discriminating if its strategy p is discriminating.

Contrite strategies and systems. A strategy p is ‘contrite’ if it cooperates whenever the focal player is in bad standing ($r_1 = B$), that is, if $p(B, B) = C$ and $p(B, G) = C$; as a string, $p = CC**$. Contrition thus captures the readiness to make amends – cooperating even with a bad co-player – when one’s own reputation is bad. A moral system (p, d) is contrite if its strategy p is contrite; among the four discriminating strategies, CDISC is the only contrite one, so a discriminating system is contrite precisely when its strategy is CDISC.

Consistent systems. A moral system (p, d) is ‘consistent’ if the social norm d judges individuals in strict accordance with the strategy p , whenever the co-player behaves as p prescribes. Formally, (p, d) is consistent if, for all $r_1, r_2 \in \{B, G\}$:

$$d(r_1, r_2, p(r_1, r_2), p(r_2, r_1)) = G, \quad (2a)$$

$$d(r_1, r_2, \neg p(r_1, r_2), p(r_2, r_1)) = B, \quad (2b)$$

where $\neg C = D$ and $\neg D = C$. Consistency is silent about what reputation to assign when the co-player itself deviates from p : conditions (2) fix eight of the social norm’s sixteen bits, and the other eight – covering exactly these deviation cases – remain free. Each strategy p therefore gives rise to exactly $2^8 = 256$ consistent social norms. In the absence of execution errors ($\epsilon = 0$), co-players – who are residents under the rare-mutant assumption – always execute the prescribed action, so these unconstrained bits are never invoked. All 256 consistent social norms for a given strategy therefore produce identical reputation dynamics, payoffs, and hence stability and

cooperation indices in this case (see SI Remark 1). A moral system that is both discriminating and consistent is a ‘consistent discriminating system’; the four discriminating strategies therefore give $4 \times 256 = 1,024$ such systems (their social norms are shown in Fig. 3a).

Evolutionary analysis

Reputation dynamics and equilibrium. We consider an infinite, well-mixed population whose members are repeatedly and randomly paired to play the game, with a single observer assessing both players after each interaction (public assessment). In a resident population implementing moral system (p, d) , let $g_{p,d}$ denote the fraction of individuals in good standing. An individual with reputation r_1 who meets a co-player with reputation r_2 is then assigned a good reputation with probability $\delta_{r_1 r_2}$, a function of the social norm d and the error rates ϵ and χ (SI Eq. (19)). Averaging over the reputations of randomly matched co-players yields a one-dimensional dynamic for $g_{p,d}$, with a unique, asymptotically stable equilibrium $g_{p,d}^*$ (SI Eqs. (20), (21)). All resident quantities below are evaluated at this reputation equilibrium.

Cooperation index. Each moral system (p, d) is characterized by a cooperation index $\kappa_{p,d}$, the probability that, at equilibrium, a randomly chosen resident cooperates in a social interaction. It thus measures how cooperative the moral system is, and is independent of the game used to model social interactions, though it depends on the error rates ϵ and χ . Concretely, $\kappa_{p,d}$ is a resident’s probability of cooperating, averaged over the reputations of two randomly matched residents and evaluated at the reputation equilibrium $g_{p,d}^*$ (SI Eq. (23)).

Expected payoffs and stability. Fix a cooperative dilemma (determined by a class $\Gamma \in \{\text{PD}, \text{SG}, \text{SH}\}$ together with specific values of its two payoff parameters), error rates (ϵ, χ) , and a social norm d . Strategy p is stable under d if, in a resident population implementing (p, d) , its expected payoff $w_{p,d}$ strictly exceeds that of every rare mutant strategy:

$$w_{p,d} > \tilde{w}_{\tilde{p},p,d} \quad \text{for all } \tilde{p} \neq p, \quad (3)$$

where $\tilde{w}_{\tilde{p},p,d}$ denotes the expected payoff of a mutant using $\tilde{p}$. If condition (3) holds, then p constitutes a strict Nash equilibrium of the induced population game under social norm d . The payoffs $w_{p,d}$ and $\tilde{w}_{\tilde{p},p,d}$ are computed at the joint reputation equilibrium. Because mutants are vanishingly rare, they interact only with residents and leave the resident reputation equilibrium $g_{p,d}^*$ unperturbed, while the mutant sub-population settles at its own unique, asymptotically

stable equilibrium fraction of good reputations $h_{\tilde{p},p,d}^*$ against that resident background (SI Eqs. (25), (26)). Each payoff is then the expected game payoff averaged over the reputations of the interacting players – resident–resident encounters for $w_{p,d}$, and mutant–resident encounters for $\tilde{w}_{\tilde{p},p,d}$ – evaluated at $(g_{p,d}^*, h_{\tilde{p},p,d}^*)$ and adjusted for execution errors (SI Eqs. (28), (30)). We determine stability conditions with a combination of analytical and numerical techniques that extend methods previously developed for the donation game and social norms up to third order^{3,5}.

Stability index. Let $\mathcal{G}_\Gamma$ denote the normalized two-dimensional payoff region corresponding to game class $\Gamma \in \{\text{PD}, \text{SG}, \text{SH}\}$. For fixed error rates (ϵ, χ) and a moral system (p, d) , define

$$\mathcal{S}_{p,d}^\Gamma(\epsilon, \chi) := \{ \pi \in \mathcal{G}_\Gamma : p \text{ is stable under } d \text{ at payoff } \pi \}, \quad (4)$$

where stability at a given payoff configuration is understood in the sense defined above. We define the stability index

$$\sigma_{p,d}^\Gamma(\epsilon, \chi) := \frac{\lambda_2(\mathcal{S}_{p,d}^\Gamma(\epsilon, \chi))}{\lambda_2(\mathcal{G}_\Gamma)}, \quad (5)$$

where λ_2 denotes the two-dimensional Lebesgue measure on the payoff parameter space. Thus $\sigma_{p,d}^\Gamma \in [0, 1]$ represents the normalized area of the payoff region within class Γ for which p is stable under d . Equivalently, $\sigma_{p,d}^\Gamma$ is the probability that p is stable under d when a payoff configuration is drawn uniformly at random from $\mathcal{G}_\Gamma$.

To compare moral systems across the three classes of cooperative dilemma, we also report the average stability index

$$\sigma_{p,d}(\epsilon, \chi) := \frac{1}{3}\sigma_{p,d}^{\text{PD}}(\epsilon, \chi) + \frac{1}{3}\sigma_{p,d}^{\text{SG}}(\epsilon, \chi) + \frac{1}{3}\sigma_{p,d}^{\text{SH}}(\epsilon, \chi). \quad (6)$$

This average weights the three dilemma classes equally, regardless of the differing areas of their payoff regions. As a robustness check, we further compute an alternative aggregate stability index in which each class Γ is weighted in proportion to the area $\lambda_2(\mathcal{G}_\Gamma)$ of its normalized payoff region (Extended Data Fig. 2); see SI Fig. 27.

Numerical evaluation. For every moral system, we test the stability condition (3) against all 15 alternative strategies. With no execution errors and in the limit of vanishing assignment errors ($\epsilon = 0, \chi \rightarrow 0$), the stability region $\mathcal{S}_{p,d}^\Gamma$ is a polygon whose vertices we obtain analytically and whose area we evaluate in closed form with the shoelace formula (SI Section 1.10). This yields the exact stability indices reported in Fig. 4a,b (computed in SI Sections 3.6 and 3.7, respectively).

At the finite error rates used for our main results ($\epsilon = \chi = 0.02$; Fig. 2), we instead estimate each stability index by Monte Carlo integration, drawing 1,000 payoff configurations uniformly at random from the normalized region $\mathcal{G}_\Gamma$ of each dilemma class and recording the fraction of configurations at which p is stable under d . Because a moral system and its ‘mirror’ – the system obtained by swapping the two reputation labels throughout – are behaviorally equivalent, we evaluate one representative per mirror class, 524,800 in total (SI Sections 4.1 and 4.2).

Righteous and universally stable systems. Our central analytical results concern several further properties of moral systems, all evaluated in a tractable regime: no execution errors ($\epsilon = 0$) and the limit of vanishing assignment errors ($\chi \rightarrow 0$). A moral system is ‘self-approving’ if almost all of its residents attain good standing, that is, if the resident reputation equilibrium approaches one, $\lim_{\chi \rightarrow 0} (g_{p,d}^*)_{\epsilon=0} = 1$ (SI Definition 8). It is ‘self-cooperative’ if its residents cooperate at near-maximal rates, that is, if the cooperation index approaches one, $\lim_{\chi \rightarrow 0} (\kappa_{p,d})_{\epsilon=0} = 1$ (SI Definition 10). It is ‘righteous’ if it is both self-approving and self-cooperative (SI Definition 11). Finally, it is ‘universally stable’ if it is stable across every cooperative dilemma we consider, that is, if its stability index approaches one in each of the three classes, $\lim_{\chi \rightarrow 0} (\sigma_{p,d}^\Gamma)_{\epsilon=0} = 1$ for all $\Gamma \in \{\text{PD}, \text{SG}, \text{SH}\}$ (SI Definition 12).

Data availability

The archived payoff samples used for the Monte Carlo stability estimates are distributed with the code (see [Code availability](#)). The computed cooperation and stability indices for all moral systems and error-rate combinations, from which the figures are generated (SI Section 4.2), are deposited at <https://doi.org/10.26180/33068042>.

Code availability

The Julia package `MoralConsistency`, which implements the reputation dynamics, fitness computations, and stability tests, together with the Python scripts that generate the figures, is openly available at <https://github.com/ghostleopold/moral-consistency-reproducibility> and permanently archived at <https://doi.org/10.5281/zenodo.21519395>.

References

1. Alexander, R. *The Biology of Moral Systems* (Aldine de Gruyter, New York, 1987).

2. Nowak, M. A. & Sigmund, K. Evolution of indirect reciprocity by image scoring. *Nature* **393**, 573–577 (1998).
3. Ohtsuki, H. & Iwasa, Y. How should we define goodness?—reputation dynamics in indirect reciprocity. *Journal of Theoretical Biology* **231**, 107–20 (2004).
4. Nowak, M. A. & Sigmund, K. Evolution of indirect reciprocity. *Nature* **437**, 1291–1298 (2005).
5. Nakamura, M. & Ohtsuki, H. Indirect reciprocity in three types of social dilemmas. *Journal of Theoretical Biology* **355**, 117–127 (2014).
6. Fehr, E. & Fischbacher, U. The nature of human altruism. *Nature* **425**, 785–791 (2003).
7. Boyd, R. & Richerson, P. J. Large-scale cooperation in small-scale foraging societies. *Evolutionary Anthropology* **31**, 175–198 (2022).
8. Trivers, R. L. The evolution of reciprocal altruism. *Quarterly Review of Biology* **46**, 35–57 (1971).
9. Boyd, R. & Richerson, P. J. The evolution of indirect reciprocity. *Social Networks* **11**, 213–236 (1989).
10. Kandori, M. Social norms and community enforcement. *The Review of Economic Studies* **59**, 63–80 (1992).
11. Sugden, R. *The Economics of Rights, Co-operation and Welfare* (Springer, 1986).
12. Ohtsuki, H. & Iwasa, Y. The leading eight: Social norms that can maintain cooperation by indirect reciprocity. *Journal of Theoretical Biology* **239**, 435–444 (2006).
13. Ohtsuki, H. & Iwasa, Y. Global analyses of evolutionary dynamics and exhaustive search for social norms that maintain cooperation by reputation. *Journal of Theoretical Biology* **244**, 518–531 (2007).
14. Pacheco, J. M., Santos, F. C. & Chalub, F. A. C. C. Stern-judging: A simple, successful norm which promotes cooperation under indirect reciprocity. *PLoS Computational Biology* **2**, 1634–1638 (2006).
15. Schmid, L., Ekbatani, F., Hilbe, C. & Chatterjee, K. Quantitative assessment can stabilize indirect reciprocity under imperfect information. *Nature Communications* **14**, 2086 (2023).
16. Okada, I. A review of theoretical studies on indirect reciprocity. *Games* **11**, 27 (2020).

17. Santos, F. P., Santos, F. C. & Pacheco, J. M. Social norm complexity and past reputations in the evolution of cooperation. *Nature* **555**, 242–245 (2018).
18. Xu, J., García, J. & Handfield, T. Cooperation with bottom-up reputation dynamics. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, 269–276 (2019).
19. Kessinger, T. A., Tarnita, C. E. & Plotkin, J. B. Evolution of norms for judging social behavior. *Proceedings of the National Academy of Sciences* **120**, e2219480120 (2023).
20. Hilbe, C., Schmid, L., Tkadlec, J., Chatterjee, K. & Nowak, M. A. Indirect reciprocity with private, noisy, and incomplete information. *Proceedings of the National Academy of Sciences* 201810565 (2018).
21. Perret, C., Krellner, M. & Han, T. A. The evolution of moral rules in a model of indirect reciprocity with private assessment. *Scientific Reports* **11**, 23581 (2021).
22. Fujimoto, Y. & Ohtsuki, H. Who is a leader in the leading eight? indirect reciprocity under private assessment. *PRX Life* **2**, 023009 (2024).
23. Michel-Mata, S. *et al.* The evolution of private reputations in information-abundant landscapes. *Nature* **634**, 883–889 (2024).
24. Peña, J. & Nöldeke, G. Cooperative dilemmas with binary actions and multiple players. *Dynamic Games and Applications* **13**, 1156–1193 (2023).
25. Skyrms, B. *The Stag Hunt and the Evolution of Social Structure* (Cambridge University Press, Cambridge, 2004).
26. Doebeli, M. & Hauert, C. Models of cooperation based on the prisoner’s dilemma and the snowdrift game. *Ecology Letters* **8**, 748–766 (2005).
27. Leimar, O. & Hammerstein, P. Evolution of cooperation through indirect reciprocity. *Proceedings of the Royal Society B* **268**, 745–753 (2001).
28. Murase, Y. & Hilbe, C. Indirect reciprocity under opinion synchronization. *Proceedings of the National Academy of Sciences* **121**, e2418364121 (2024).
29. Chalub, F. A. C. C., Santos, F. C. & Pacheco, J. M. The evolution of norms. *Journal of Theoretical Biology* **241**, 233–240 (2006).

30. Boehm, C. *Blood Revenge: The Enactment and Management of Conflict in Montenegro and Other Tribal Societies* (University of Pennsylvania Press, Philadelphia, 1984).
31. Nisbett, R. E. & Cohen, D. *Culture of Honor: The Psychology of Violence in the South* (Westview Press, Boulder, 1996).
32. Thrasher, J. & Handfield, T. Honor and violence. *Human Nature* **29**, 371–389 (2018).
33. Milinski, M., Semmann, D., Bakker, T. C. & Krambeck, H.-J. Cooperation through indirect reciprocity: image scoring or standing strategy? *Proceedings of the Royal Society B* **268**, 2495–2501 (2001).
34. Milinski, M., Semmann, D. & Krambeck, H. Donors to charity gain in both indirect reciprocity and political reputation. *Proceedings of the Royal Society B* **269**, 881–883 (2002).
35. Wedekind, C. & Braithwaite, V. A. The long-term benefits of human generosity in indirect reciprocity. *Current Biology* **12**, 1012–1015 (2002).
36. Romano, A. *et al.* Reputation and socio-ecology in humans. *Philosophical Transactions of the Royal Society B* **376**, 20200295 (2021).
37. Seinen, I. & Schram, A. Social status and group norms: Indirect reciprocity in a repeated helping experiment. *European Economic Review* **50**, 581–602 (2006).
38. Swakman, V., Molleman, L., Ule, A. & Egas, M. Reputation-based cooperation: empirical evidence for behavioral strategies. *Evolution and Human Behavior* **37**, 230–235 (2016).
39. Okada, I., Yamamoto, H., Sato, Y., Uchida, S. & Sasaki, T. Experimental evidence of selective inattention in reputation-based cooperation. *Scientific Reports* **8**, 1–7 (2018).
40. Gong, B. & Yang, C.-L. Cooperation through indirect reciprocity: The impact of higher-order history. *Games and Economic Behavior* **118**, 316–341 (2019).
41. Colnaghi, M., Santos, F. P., Van Lange, P. A. M. & Balliet, D. Adaptations to infer fitness interdependence promote the evolution of cooperation. *Proceedings of the National Academy of Sciences* **120**, e2312242120 (2023).
42. Radzvilavicius, A. L., Stewart, A. J. & Plotkin, J. B. Evolution of empathetic moral evaluation. *eLife* **8**, e44269 (2019).
43. Schmid, L., Shati, P., Hilbe, C. & Chatterjee, K. The evolution of indirect reciprocity under action and assessment generosity. *Scientific Reports* **11**, 17443 (2021).

Acknowledgements

We are grateful to Fernando P. Santos for valuable discussions during the early stages of this project, and to Jean-François Bonnefon and Catherine Molho for feedback on an earlier draft. J.P. is grateful to the Institute for Advanced Study (IAS) of the University of Amsterdam for hosting the sabbatical research fellowship during which this project began. J.P. acknowledges financial support from the French National Research Agency (ANR-22-CE26-0014-01 and ANR-17-EURE-0010). T.H.'s work on this project was supported by the Australian Research Council, FT180100067.

Author contributions

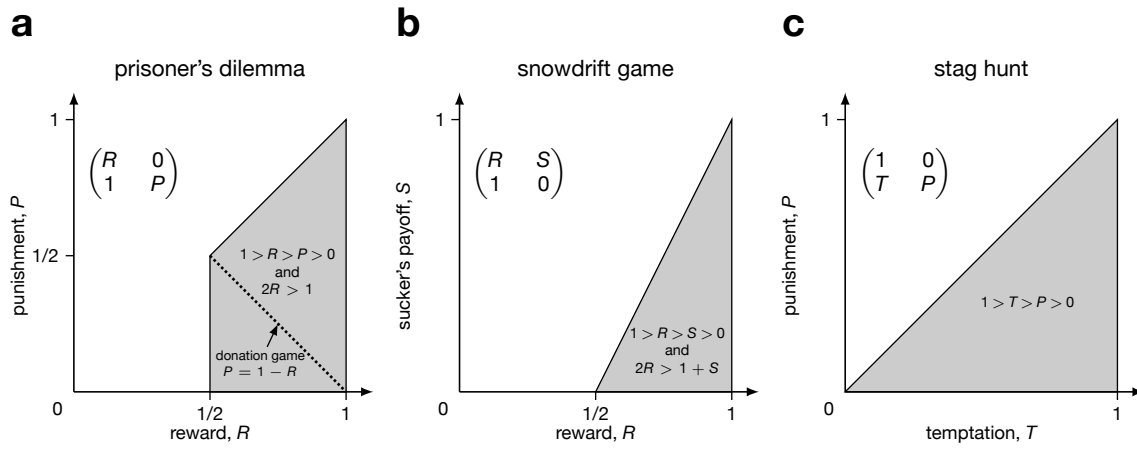
J.P., J.G., and T.H. conceived and designed the study. J.P. derived the analytical results and wrote the Supplementary Information. J.G. developed the Julia package and performed the numerical analyses. T.H. produced the figures and numerical robustness analyses. J.P. and T.H. wrote the first draft of the manuscript. All authors revised the manuscript.

Competing interests

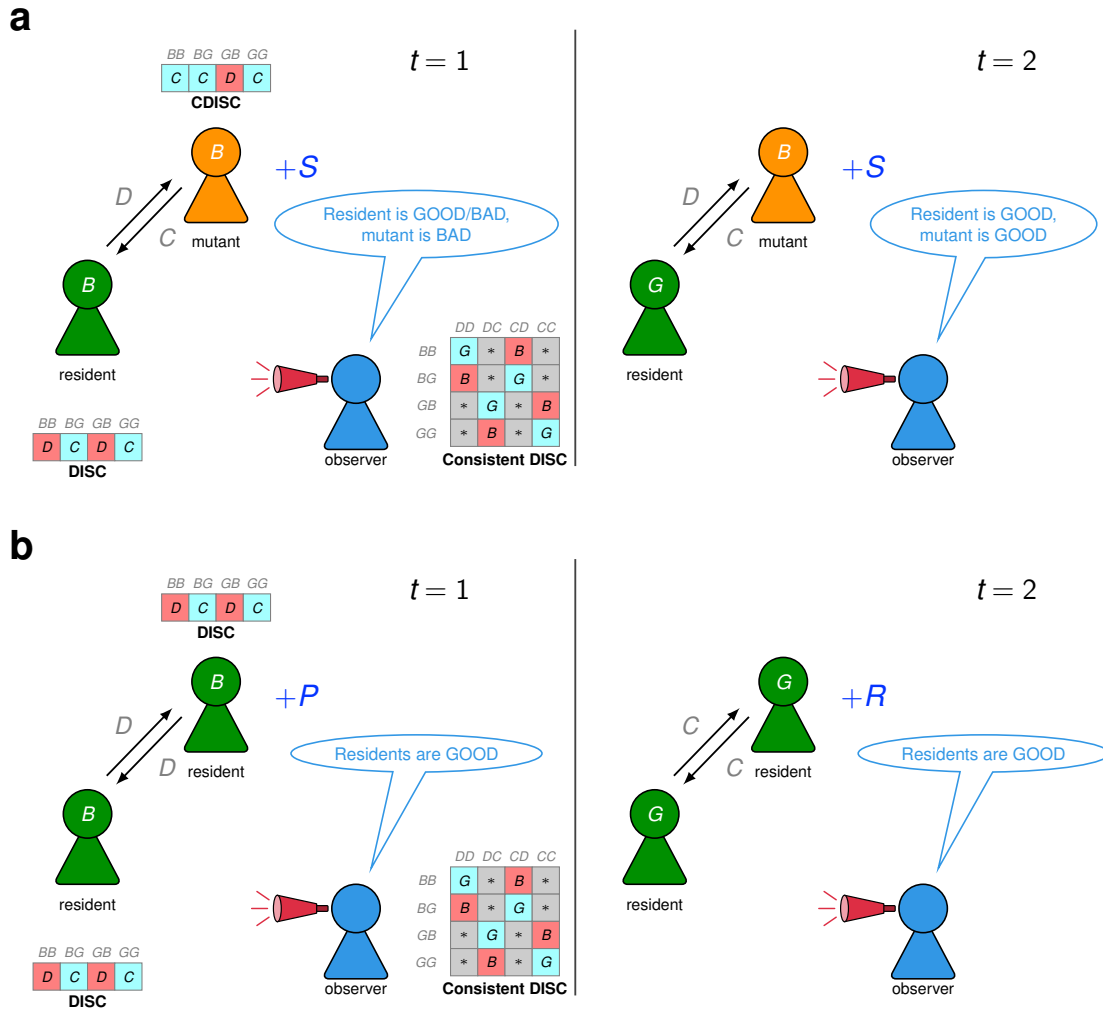
The authors declare no competing interests.

Additional information

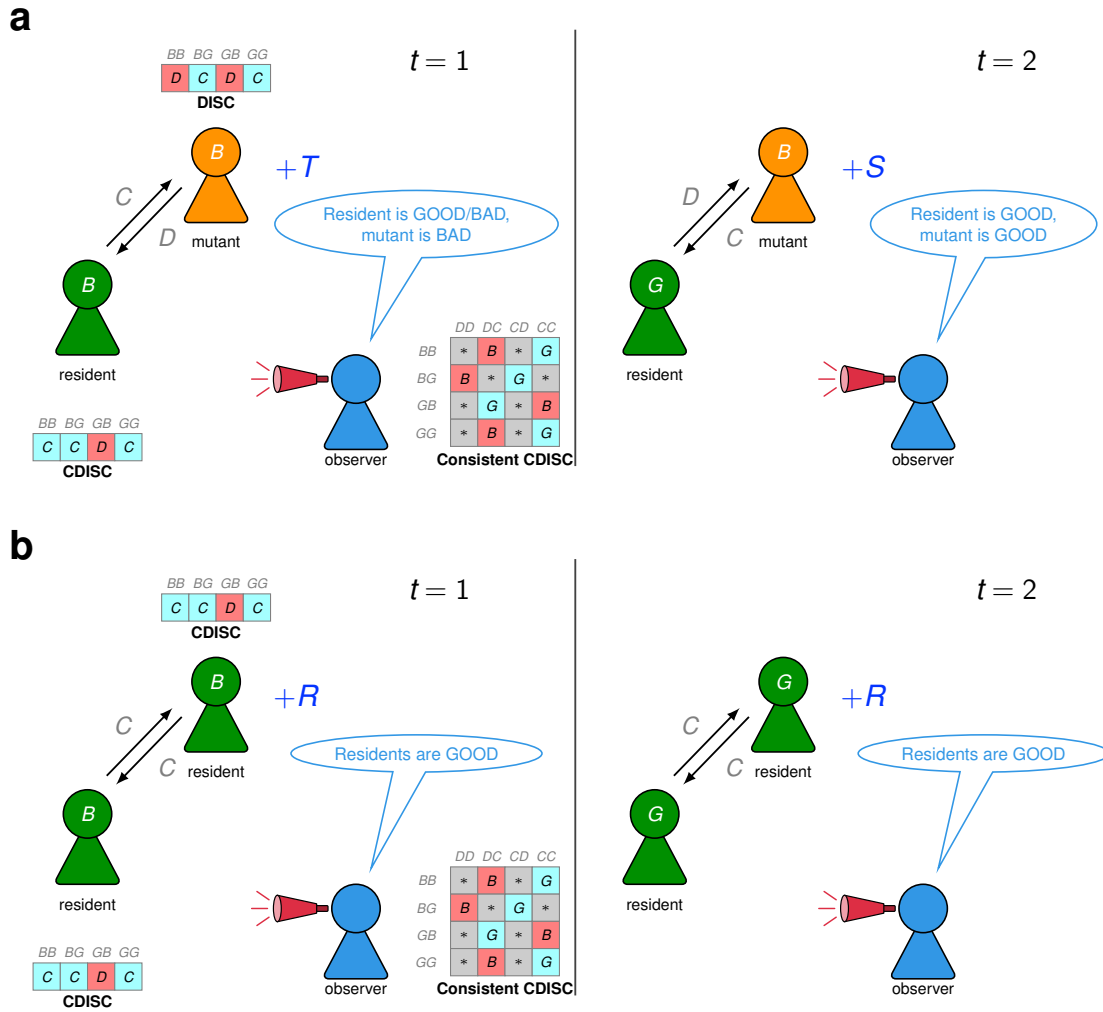
Correspondence and requests for materials should be addressed to Jorge Peña (jorge.pena@tse-fr.eu).



Extended Data Fig. 2: **The space of 2×2 cooperative dilemmas.** We consider the three classes of 2×2 cooperative dilemma, shown after the min-max normalization of payoffs to $[0, 1]$ (Methods); throughout, we impose $2R > T + S$ so that mutual cooperation is socially optimal. **a**, Prisoner's dilemmas (PD), with payoff ranking $1 > R > P > 0$; the efficiency condition adds the constraint $2R > 1$. The donation game, used in most previous studies of indirect reciprocity, is the special case with $P = 1 - R$ (dotted line). **b**, Snowdrift games (SG), with payoff ranking $1 > R > S > 0$; the efficiency condition adds the constraint $2R > 1 + S$. **c**, Stag hunts (SH), with payoff ranking $1 > T > P > 0$; the efficiency condition holds automatically.



Extended Data Fig. 3: **How CDISC can invade a consistent DISC system.** CDISC and DISC differ only in the rare, error-induced state where both players have bad reputations. Pawns carry their reputations (G or B) and follow the strategies shown beside them (Fig. 1b); grey letters give actions, and blue annotations the focal player's payoffs. Each round, an observer (*blue pawn*) judges both players by a DISC-consistent social norm (* marks entries unconstrained by consistency; Fig. 3a) and broadcasts the verdicts. **a**, Trajectory of a mutant CDISC (*orange*) starting in bad standing. At $t = 1$, the mutant meets a resident also in bad standing and contritely cooperates, earning S ; because cooperating with a bad partner deviates from DISC's prescription, it is judged bad (the resident's verdict is unconstrained, hence 'GOOD/BAD'). At $t = 2$, it meets another resident, almost surely in good standing, and cooperates, matching DISC's prescription; it is judged good but its co-player defects, so it earns S again. Its reputation is restored only after two rounds, for a total payoff of $2S$. **b**, Trajectory of a resident DISC in the analogous situation. At $t = 1$, it defects, as DISC prescribes, earns P , and is judged good; at $t = 2$, in good standing like almost all residents, it cooperates and earns R , for a total of $P + R$. CDISC thus gains a net advantage, and can invade, whenever $2S > R + P$, which only snowdrift games satisfy.



Extended Data Fig. 4: **Why DISC cannot invade a consistent CDISC system.** The converse of Extended Data Fig. 3: a rare DISC mutant among CDISC residents, judged by a social norm consistent with CDISC; notation as in Extended Data Fig. 3. **a**, Trajectory of a mutant DISC (orange) starting from a bad reputation. At $t = 1$, the mutant meets a resident who is also in bad standing and defects on its contritely cooperating partner; it earns the temptation payoff T but, because defecting on a bad partner deviates from what CDISC prescribes, is judged bad (the resident's verdict again falls on an unconstrained entry, hence 'GOOD/BAD'). At $t = 2$, the mutant meets another resident, almost surely in good standing, and cooperates, as both strategies prescribe; it is judged good, but its co-player defects and it earns only S . Its reputation is restored after two rounds, over which it accumulates $T + S$. **b**, Trajectory of a resident CDISC in the analogous situation. At $t = 1$, it cooperates with its bad partner, as CDISC prescribes, earns R , and is immediately judged good; at $t = 2$ it cooperates again and earns R , for a total of $2R$. Because we require mutual cooperation to be efficient, $2R > T + S$ holds in every dilemma we study: DISC never gains an advantage and cannot invade a consistent CDISC system.

Cooperation thrives when moral judgment matches moral conduct: Supplementary Information

Jorge Peña^{1,*}, Julian Garcia², and Toby Handfield³

¹Department of Social and Behavioral Sciences, Toulouse School of Economics, Toulouse, France

²Department of Data Science and AI, Faculty of Information Technology, Monash University, Melbourne, Australia

³School of Philosophical, Historical and International Studies, Monash University, Melbourne, Australia

*Corresponding author: jorge.pena@tse-fr.eu

Contents

1	Model	3
1.1	Labeling conventions for actions and reputations	3
1.2	Payoff function	3
1.3	Cooperative dilemmas and payoff spaces	4
1.4	Strategies, assessment rules, and shorthand notations	5
1.5	Counting assessment rules by the arguments they use	5
1.6	Resident reputation dynamics and equilibrium	7
1.7	Cooperation index	7
1.8	Mutant reputation dynamics and equilibrium	8
1.9	Equilibrium payoffs and evolutionary stability	8
1.10	Stability indices	9
1.10.1	Areas of the payoff regions	10
1.10.2	Stability indices for two recurring binding conditions	10
1.10.3	Weighted average stability indices	11
2	Definitions and preliminary results	12
2.1	Conditional, discriminating, and contrite strategies and systems	13
2.2	Self-improving, self-cooperative, righteous, and universally stable systems	15
2.3	Consistent systems and consistent discriminating systems	17
2.3.1	Consistent systems	17
2.3.2	Consistent discriminating systems	17
2.4	The leading eight	18
2.5	Mirror systems	21
3	Analytical results	23
3.1	Characterization of self-improving systems	23
3.2	Characterization of self-cooperative and righteous systems	27
3.3	Righteousness and reputational dominance of consistent systems	29
3.4	Righteousness of the leading eight	33

3.5	Universal stability within the set of righteous systems	33
3.5.1	Preliminaries	33
3.5.2	Invasion attempts by ALLD	36
3.5.3	Invasion attempts by ALLC	40
3.5.4	Invasion attempts by a discriminating mutant	44
3.5.5	The universal stability theorem	47
3.6	Stability analysis of consistent discriminating systems	48
3.6.1	Consistent discriminating systems with CDISC as action rule	48
3.6.2	Consistent discriminating systems with DISC as action rule	49
3.6.3	Consistent discriminating systems with ODISC as action rule	50
3.6.4	Consistent discriminating systems with PDISC as action rule	51
3.6.5	Summary: consistent discriminating systems	52
3.7	Stability analysis of the leading eight	52
3.7.1	L1 (CDISC, Standing)	53
3.7.2	L2 (CDISC, Consistent Standing)	54
3.7.3	L3 (DISC, Simple Standing)	55
3.7.4	L4 (DISC, SS-SJ)	56
3.7.5	L5 (DISC, SJ-SS)	57
3.7.6	L6 (DISC, Stern Judging)	58
3.7.7	L7 (DISC, Staying)	58
3.7.8	L8 (DISC, Judging)	60
3.7.9	Summary: the leading eight	61
4	Numerical results	61
4.1	Mirror classes and representatives	62
4.2	Computation of the cooperation and stability indices	64
4.3	Alternative error rates	65
4.4	Disaggregated stability indices	84
4.5	Area-weighted stability index	88

1 Model

This section provides a detailed description of the model used in the main text. We consider an infinite, well-mixed population whose members interact in pairwise encounters, each modeled as a symmetric two-player game. Each game belongs to one of three classes of cooperative dilemmas (prisoner’s dilemmas, snowdrift games, or stag hunts); throughout we restrict attention to those for which mutual cooperation is the social optimum, an efficiency condition on the payoffs that we make precise below. Individuals choose their actions according to a strategy (or action rule) that conditions on the binary reputations of the focal player and their co-player, while an observer updates both players’ reputations according to an assessment rule (or social norm) that conditions additionally on the two actions taken. Sections 1.1–1.3 establish labeling conventions and define the payoff function and the three classes of cooperative dilemmas. Section 1.4 introduces notation for strategies and assessment rules, and Section 1.5 counts assessment rules by the arguments they use, situating the space of social norms we explore relative to those studied in previous work. Section 1.6 describes the reputation dynamics of a resident population, from which Section 1.7 derives the cooperation index; Section 1.8 extends the reputation dynamics to rare mutants, and Section 1.9 derives the equilibrium payoffs and defines evolutionary stability. Finally, Section 1.10 defines the stability indices we use to assess the evolutionary robustness of moral systems across the full space of cooperative dilemmas.

1.1 Labeling conventions for actions and reputations

Actions and reputations are binary. In each interaction, an individual either ‘cooperates’ (C) or ‘defects’ (D); after the interaction, an observer may assign the individual a ‘good’ (G) or a ‘bad’ (B) reputation. For notational convenience, we identify cooperation and a good reputation with 1, and defection and a bad reputation with 0:

$$C = G = 1, \text{ and} \quad (1a)$$

$$D = B = 0. \quad (1b)$$

We also use the wildcard symbol $*$ to denote a placeholder that may represent either 0 (D or B) or 1 (C or G).

1.2 Payoff function

Social interactions are modeled as symmetric 2×2 games. A focal player choosing action $a_1 \in \{D, C\}$ against a co-player choosing action $a_2 \in \{D, C\}$ receives a payoff $v(a_1, a_2)$, where $v : \{D, C\}^2 \rightarrow \mathbb{R}$. Writing the focal player’s action as the row and the co-player’s as the column, this payoff function takes the matrix form

$$\begin{array}{cc} & \begin{array}{cc} C & D \end{array} \\ \begin{array}{c} C \\ D \end{array} & \left(\begin{array}{cc} v(C, C) & v(C, D) \\ v(D, C) & v(D, D) \end{array} \right). \end{array} \quad (2)$$

Identifying

$$v(C, C) = R, \quad v(C, D) = S, \quad v(D, C) = T, \quad \text{and} \quad v(D, D) = P,$$

we recover the formulation given in the main text (Fig. 1a; see also [Methods](#)).

Throughout, we also use the shorthand notation:

$$v_{a_1 a_2} = v(a_1, a_2).$$

1.3 Cooperative dilemmas and payoff spaces

We focus on the generic 2×2 games that constitute cooperative dilemmas in which mutual cooperation is the social optimum¹. This requires that all of the following inequalities hold:

$$R > P, \quad (3)$$

$$R > S, \quad (4)$$

$$T > P, \quad (5)$$

$$2R > T + S, \quad (6)$$

and that either

$$T > R \quad (7)$$

or

$$P > S \quad (8)$$

also holds. Inequality (3) means that mutual cooperation is preferred over mutual defection. Inequalities (4) and (5) imply that each player prefers the co-player to cooperate rather than defect. Inequality (6) ensures that two cooperators earn a higher total payoff ($2R$) than a pair in which one cooperates and the other defects ($T + S$); together with (3), this makes mutual cooperation the social optimum. Inequalities (7) and (8) indicate that there is some temptation for players to defect; this is a necessary condition for the game to be a social dilemma¹.

Three classes of games fulfill these requirements¹: prisoner's dilemmas (PD), characterized by

$$T > R > P > S \text{ and } 2R > T + S; \quad (9)$$

snowdrift games (SG), characterized by

$$T > R > S > P \text{ and } 2R > T + S; \quad (10)$$

and stag hunts (SH), characterized by

$$R > T > P > S, \quad (11)$$

which trivially satisfy (6). Note that all three classes satisfy

$$T > S. \quad (12)$$

Normalizing the payoffs by a min-max transformation ([Methods](#)), so that the largest payoff equals one and the smallest equals zero, we obtain

$$1 > R > P > 0 \text{ and } 2R > 1 \quad (13)$$

for prisoner's dilemmas,

$$1 > R > S > 0 \text{ and } 2R > 1 + S \quad (14)$$

for snowdrift games, and

$$1 > T > P > 0 \quad (15)$$

for stag hunts. This fixes two of the four payoffs, leaving each class with two free parameters (see [Extended Data Fig. 2](#) of the main text).

1.4 Strategies, assessment rules, and shorthand notations

Consider a population where observers judge players according to assessment rule d , resident players adopt strategy p , and a small fraction of mutant players adopt strategy $\tilde{p}$. As specified in the [Methods](#) of the main text, the strategies p and $\tilde{p}$ are Boolean functions $p, \tilde{p} : \{B, G\}^2 \rightarrow \{D, C\}$ prescribing an action for each combination of the reputations of the focal player and their co-player. Likewise, the assessment rule is a Boolean function $d : \{B, G\}^2 \times \{D, C\}^2 \rightarrow \{B, G\}$ prescribing a reputation for each combination of the reputations and actions of the focal player and their co-player.

Consider a focal player who interacts with a co-player and is judged by an observer. Let $r_1 \in \{B, G\}$ denote the reputation of the focal player and $r_2 \in \{B, G\}$ that of the co-player. The focal player's strategy prescribes action $p(r_1, r_2)$ if they are a resident and $\tilde{p}(r_1, r_2)$ if they are a mutant. Once the focal player has taken an action $a_1 \in \{D, C\}$ against a co-player who has played $a_2 \in \{D, C\}$, the assessment rule prescribes a new reputation $d(r_1, r_2, a_1, a_2)$ for the focal player.

It is convenient to introduce shorthand notation for the actions and reputations prescribed in different situations. We use the following conventions throughout:

$$p_{r_1 r_2} := p(r_1, r_2), \quad (16a)$$

$$\tilde{p}_{r_1 r_2} := \tilde{p}(r_1, r_2), \quad (16b)$$

$$d_{r_1 r_2 a_1 a_2} := d(r_1, r_2, a_1, a_2), \quad (16c)$$

$$d_{r_1 r_2} := d(r_1, r_2, p_{r_1 r_2}, p_{r_2 r_1}), \quad (16d)$$

$$\tilde{d}_{r_1 r_2} := d(r_1, r_2, \tilde{p}_{r_1 r_2}, p_{r_2 r_1}), \quad (16e)$$

$$\bar{d}_{r_1 r_2} := d(r_1, r_2, \neg p_{r_1 r_2}, p_{r_2 r_1}), \quad (16f)$$

where $\neg$ stands for the logical negation (with our labeling convention (1), $\neg C = D$ and $\neg D = C$). In expressions (16d)–(16f) the co-player is a resident who behaves as prescribed, playing $p(r_2, r_1)$; the three differ only in the focal player's action. Expression (16d) is the new reputation prescribed to a resident focal player, who plays $p(r_1, r_2)$, and (16e) that prescribed to a mutant focal player, who plays $\tilde{p}(r_1, r_2)$ – both in the absence of execution errors. Expression (16f) is the new reputation prescribed to a focal player (resident or mutant) whose action deviates from the resident prescription, playing $\neg p(r_1, r_2)$.

1.5 Counting assessment rules by the arguments they use

[Extended Data Fig. 1](#) of the main text situates the space of assessment rules we study relative to the spaces explored in previous work. Here we define the sets appearing there and derive their cardinalities.

An assessment rule is a Boolean function of the four binary arguments in $\mathcal{V} = \{r_1, r_2, a_1, a_2\}$. For a subset $\mathcal{U} \subseteq \mathcal{V}$, let $\mathcal{T}_{\mathcal{U}}$ denote the set of assessment rules of type $\mathcal{U}$: those that can be written as functions of the arguments in $\mathcal{U}$ alone. There are $|\mathcal{T}_{\mathcal{U}}| = 2^{2^{|\mathcal{U}|}}$ rules of type $\mathcal{U}$, and $\mathcal{T}_{\mathcal{U}} \cap \mathcal{T}_{\mathcal{U}'} = \mathcal{T}_{\mathcal{U} \cap \mathcal{U}'}$, since a rule that ignores the arguments outside $\mathcal{U}$ and the arguments outside $\mathcal{U}'$ ignores all arguments outside $\mathcal{U} \cap \mathcal{U}'$. For $k = 0, 1, \dots, 4$, let

$$O_k := \bigcup_{\mathcal{U} \subseteq \mathcal{V}, |\mathcal{U}|=k} \mathcal{T}_{\mathcal{U}}$$

denote the set of assessment rules of order at most k ; in particular, O_4 is the universe of $2^{2^4} = 2^{16} = 65,536$ assessment rules studied in this paper.

A rule is said to depend *essentially* on the subset $\mathcal{U}$ if it is of type $\mathcal{U}$ but not of type $\mathcal{U}'$ for any proper subset $\mathcal{U}' \subset \mathcal{U}$. The following proposition counts assessment rules by their essential arguments and, with them, the sets O_k .

Proposition 1 (Number of assessment rules of each order). *For $k = 0, 1, \dots, 4$, the number of assessment rules that depend essentially on a given subset $\mathcal{U} \subseteq \mathcal{V}$ with k elements depends only on k . Denoting this number by N_k , it satisfies the recursion*

$$N_k = 2^{2^k} - \sum_{j < k} \binom{k}{j} N_j, \quad (17)$$

giving $N_0 = 2$, $N_1 = 2$, $N_2 = 10$, $N_3 = 218$, and $N_4 = 64,594$. Moreover, the sets of assessment rules of order at most k have cardinalities

$$|O_k| = \sum_{j \leq k} \binom{4}{j} N_j, \quad (18)$$

giving $|O_0| = 2$, $|O_1| = 10$, $|O_2| = 70$, $|O_3| = 942$, and $|O_4| = 2^{16} = 65,536$.

Proof. Since type classes are closed under intersection, each rule is of a unique minimal type, that is, it depends essentially on exactly one subset of $\mathcal{V}$. Consequently, the $2^{2^{|\mathcal{U}|}}$ rules of type $\mathcal{U}$ partition according to the subset $\mathcal{U}' \subseteq \mathcal{U}$ on which they depend essentially, so the number $E(\mathcal{U})$ of rules depending essentially on $\mathcal{U}$ satisfies $E(\mathcal{U}) = 2^{2^{|\mathcal{U}|}} - \sum_{\mathcal{U}' \subsetneq \mathcal{U}} E(\mathcal{U}')$. By induction on $|\mathcal{U}|$, starting from $E(\emptyset) = 2$ (the two constant rules, which assign G to everybody or B to everybody), $E(\mathcal{U})$ depends only on $k = |\mathcal{U}|$, and grouping the proper subsets of $\mathcal{U}$ by their cardinalities yields the recursion (17). Finally, a rule belongs to O_k if and only if the subset on which it depends essentially has at most k elements; summing N_j over the $\binom{4}{j}$ subsets of $\mathcal{V}$ with $j \leq k$ elements yields Eq. (18). $\square$

The spaces explored by previous literature are $\mathcal{T}_{\{a_1, r_1, r_2\}}$, the $2^{2^3} = 256$ third-order rules of type $a_1 r_1 r_2$ ², and $\mathcal{T}_{\{a_1, a_2, r_2\}}$, the 256 third-order rules of type $a_1 a_2 r_2$ ³. By the intersection property above, a rule belongs to both spaces if and only if it is of type $a_1 r_2$. The two spaces therefore share $|\mathcal{T}_{\{a_1, r_2\}}| = 2^{2^2} = 16$ rules, and their union contains $256 + 256 - 16 = 496$ rules.

Consider first the $|O_2| = 70$ rules of order at most two (Proposition 1). Such a rule lies in $\mathcal{T}_{\{a_1, r_1, r_2\}}$ whenever its essential arguments avoid a_2 , and in $\mathcal{T}_{\{a_1, a_2, r_2\}}$ whenever they avoid r_1 . The rules of order at most two lying outside both previously studied spaces are therefore exactly the rules that depend essentially on the pair $a_2 r_1$, which number $N_2 = 10$. The rules of order at most two belonging to $\mathcal{T}_{\{a_1, r_1, r_2\}}$ are those whose essential arguments form one of the $1 + 3 + 3$ subsets of $\{a_1, r_1, r_2\}$ with at most two elements; they number $N_0 + 3N_1 + 3N_2 = 38$. Of these, the 16 rules of type $a_1 r_2$ are shared with $\mathcal{T}_{\{a_1, a_2, r_2\}}$, leaving $38 - 16 = 22$ exclusive ones; by symmetry, the same counts hold for $\mathcal{T}_{\{a_1, a_2, r_2\}}$.

Consider next the rules of essential order three. Those depending essentially on all of a_1, r_1, r_2 number $N_3 = 218$ and complete the count $38 + 218 = 256$ of $\mathcal{T}_{\{a_1, r_1, r_2\}}$; symmetrically for $\mathcal{T}_{\{a_1, a_2, r_2\}}$. The two remaining triples define the type classes $\mathcal{T}_{\{a_1, a_2, r_1\}}$ and $\mathcal{T}_{\{a_2, r_1, r_2\}}$, which contain the pair $a_2 r_1$ and were not previously explored. Their $2N_3 = 436$ essentially third-order rules, together with the 10 rules that depend essentially on the pair $a_2 r_1$, account for all $|O_3| - 496 = 446$ rules of order at most three lying outside the previously studied spaces. The remaining $N_4 = 64,594$ rules of the universe O_4 are of essential order four.

All of these counts were additionally verified by exhaustive enumeration of the 2^{16} assessment rules. In [Extended Data Fig. 1](#), the area of every region is proportional, within each panel, to the number of assessment rules it contains.

1.6 Resident reputation dynamics and equilibrium

Throughout, we adopt a public-assessment model of reputation^{2,4}: a single observer witnesses each interaction and assigns each of the two players a new reputation, and these assessments are then shared with the whole population (for instance, through gossip or a central authority), so that every individual holds the same view of any given player. This shared, population-wide agreement is what allows the population's state to be summarized by a single variable – the fraction of individuals in good standing – whose dynamics we now derive.

Consider a resident population implementing moral system (p, d) : all players behave according to strategy p and observers judge according to assessment rule d . Within this population, take a focal player with reputation $r_1 \in \{B, G\}$ playing a game against a co-player with reputation $r_2 \in \{B, G\}$; the resident strategy then prescribes that the focal player play $p_{r_1 r_2}$ and the co-player play $p_{r_2 r_1}$. Given our labeling convention (1), the probability that the focal player is assigned a good reputation can be written as

$$\begin{aligned} \delta_{r_1 r_2} = & \chi + (1 - 2\chi) \left\{ (1 - \epsilon)^2 d_{r_1 r_2} \right. \\ & + \epsilon(1 - \epsilon) \left[d(r_1, r_2, D, p_{r_2 r_1}) + d(r_1, r_2, p_{r_1 r_2}, D) \right] \\ & \left. + \epsilon^2 d(r_1, r_2, D, D) \right\}, \end{aligned} \quad (19)$$

where $\chi \in (0, 1/2)$ is the probability that the observer makes an assignment error², and $\epsilon \in [0, 1/2)$ is the probability that a player makes an execution error².

Let $g_{p,d}$ denote the proportion of individuals with a good reputation in the resident population. Following the same assumptions and arguments as in previous studies^{2,3}, $g_{p,d}$ changes according to the continuous-time dynamics

$$\dot{g}_{p,d} = \delta_{GG} g_{p,d}^2 + (\delta_{BG} + \delta_{GB}) g_{p,d} (1 - g_{p,d}) + \delta_{BB} (1 - g_{p,d})^2 - g_{p,d}. \quad (20)$$

Setting the right-hand side to zero and solving for $g_{p,d}$ yields the (unique) resident reputation equilibrium (ref. 3, Eq. 3)

$$g_{p,d}^* = \begin{cases} \frac{-b - \sqrt{b^2 - 4ac}}{2a} & \text{if } a \neq 0 \\ -\frac{c}{b} & \text{if } a = 0 \end{cases}, \quad (21)$$

where

$$a = \delta_{BB} + \delta_{GG} - \delta_{BG} - \delta_{GB}, \quad (22a)$$

$$b = \delta_{BG} + \delta_{GB} - 2\delta_{BB} - 1, \quad (22b)$$

$$c = \delta_{BB}. \quad (22c)$$

This equilibrium is asymptotically stable^{2,3}.

1.7 Cooperation index

We define the cooperation index $\kappa_{p,d}$ of moral system (p, d) as the probability that an individual in a population implementing (p, d) cooperates at equilibrium. This probability depends on the execution and assignment error rates ϵ and χ , and is given by

$$\kappa_{p,d} = (1 - \epsilon) \left[g^2 p_{GG} + g(1 - g) (p_{GB} + p_{BG}) + (1 - g)^2 p_{BB} \right]_{g=g_{p,d}^*}. \quad (23)$$

1.8 Mutant reputation dynamics and equilibrium

Suppose the resident population has settled at the reputation equilibrium $g_{p,d}^*$ (21), and that a vanishingly small fraction of individuals (‘mutants’) now switch to a strategy $\tilde{p} \neq p$. Since mutants are rare, they interact only with residents. A focal mutant with reputation r_1 interacting with a resident co-player with reputation r_2 is assigned a good reputation with probability

$$\tilde{\delta}_{r_1 r_2} = \chi + (1 - 2\chi) \left\{ (1 - \epsilon)^2 \tilde{d}_{r_1 r_2} + \epsilon(1 - \epsilon) [d(r_1, r_2, D, \mathbf{p}_{r_2 r_1}) + d(r_1, r_2, \tilde{\mathbf{p}}_{r_1 r_2}, D)] + \epsilon^2 d(r_1, r_2, D, D) \right\}. \quad (24)$$

The proportion $h_{\tilde{p},p,d}$ of mutants with a good reputation then changes according to the continuous-time dynamics

$$\begin{aligned} \dot{h}_{\tilde{p},p,d} = & h_{\tilde{p},p,d} g_{p,d}^* \tilde{\delta}_{GG} + h_{\tilde{p},p,d} (1 - g_{p,d}^*) \tilde{\delta}_{GB} \\ & + (1 - h_{\tilde{p},p,d}) g_{p,d}^* \tilde{\delta}_{BG} + (1 - h_{\tilde{p},p,d}) (1 - g_{p,d}^*) \tilde{\delta}_{BB} - h_{\tilde{p},p,d}. \end{aligned} \quad (25)$$

Solving for the (unique) mutant reputation equilibrium (cf. ref. 3, Eq. 8) yields

$$h_{\tilde{p},p,d}^* = \frac{\tilde{\delta}_{BB} + (\tilde{\delta}_{BG} - \tilde{\delta}_{BB}) g_{p,d}^*}{1 + \tilde{\delta}_{BB} - \tilde{\delta}_{GB} + (\tilde{\delta}_{BG} + \tilde{\delta}_{GB} - \tilde{\delta}_{BB} - \tilde{\delta}_{GG}) g_{p,d}^*}, \quad (26)$$

which is asymptotically stable^{2,3}.

1.9 Equilibrium payoffs and evolutionary stability

Consider a polymorphic population composed of residents playing strategy p and vanishingly rare mutants playing strategy $\tilde{p}$, settled at the reputation equilibrium $(g_{p,d}^*, h_{\tilde{p},p,d}^*)$; that is, the proportion of good residents is $g_{p,d}^*$ and the proportion of good mutants is $h_{\tilde{p},p,d}^*$.

Since mutants are rare, the expected payoff to a resident depends only on their interactions with other residents. Taking into account execution errors occurring with probability ϵ , the expected payoff of a resident with reputation r_1 playing against a resident with reputation r_2 is given by

$$\omega_{r_1 r_2} = (1 - \epsilon)^2 v(\mathbf{p}_{r_1 r_2}, \mathbf{p}_{r_2 r_1}) + \epsilon(1 - \epsilon) [v(\mathbf{p}_{r_1 r_2}, D) + v(D, \mathbf{p}_{r_2 r_1})] + \epsilon^2 v_{DD}. \quad (27)$$

Using this expression, the expected payoff to a resident at the resident reputation equilibrium $g_{p,d}^*$ can then be written as

$$w_{p,d} = [(1 - g)^2 \omega_{BB} + g(1 - g)(\omega_{BG} + \omega_{GB}) + g^2 \omega_{GG}]_{g=g_{p,d}^*}. \quad (28)$$

Again, since mutants are rare, the expected payoff to a mutant depends only on their interactions with residents. The expected payoff of a mutant with reputation r_1 playing against a resident with reputation r_2 is thus given by

$$\tilde{\omega}_{r_1 r_2} = (1 - \epsilon)^2 v(\tilde{\mathbf{p}}_{r_1 r_2}, \mathbf{p}_{r_2 r_1}) + \epsilon(1 - \epsilon) [v(\tilde{\mathbf{p}}_{r_1 r_2}, D) + v(D, \mathbf{p}_{r_2 r_1})] + \epsilon^2 v_{DD}. \quad (29)$$

Using this expression, the expected payoff to a mutant at the reputation equilibrium $(g_{p,d}^*, h_{\tilde{p},p,d}^*)$ can be written as

$$\tilde{w}_{\tilde{p},p,d} = [(1 - h)(1 - g)\tilde{\omega}_{BB} + (1 - h)g\tilde{\omega}_{BG} + h(1 - g)\tilde{\omega}_{GB} + hg\tilde{\omega}_{GG}]_{g=g_{p,d}^*, h=h_{\tilde{p},p,d}^*}. \quad (30)$$

The evolutionary stability of strategy p under assessment rule d can be determined once the resident payoff and the payoffs of all possible mutant strategies have been calculated from (28) and (30). Under our solution concept (strict Nash equilibrium), p is stable under d if and only if

$$w_{p,d} > \tilde{w}_{\tilde{p},p,d} \quad \text{for all } \tilde{p} \in \mathcal{P} \setminus \{p\} \quad (31)$$

holds, where $\mathcal{P}$ is the set of all strategies we consider (see also Eq. (3) in the main text). For brevity, we henceforth say that (p, d) is stable to mean that p is stable under d .

1.10 Stability indices

Let $\mathcal{G}_\Gamma$ denote the normalized two-dimensional payoff region corresponding to game class $\Gamma \in \{\text{PD}, \text{SG}, \text{SH}\}$ (see Eqs. (13)–(15)), each point $\pi \in \mathcal{G}_\Gamma$ representing a normalized payoff configuration. For fixed error rates (ϵ, χ) and a moral system (p, d) , define

$$\mathcal{S}_{p,d}^\Gamma(\epsilon, \chi) := \{ \pi \in \mathcal{G}_\Gamma : p \text{ is stable under } d \text{ at payoff } \pi \},$$

where stability at a given payoff configuration is understood in the sense of condition (31).

The stability index of p under d in class Γ is

$$\sigma_{p,d}^\Gamma(\epsilon, \chi) := \frac{\lambda_2(\mathcal{S}_{p,d}^\Gamma(\epsilon, \chi))}{\lambda_2(\mathcal{G}_\Gamma)},$$

where λ_2 denotes the two-dimensional Lebesgue measure on the payoff parameter space. Thus $\sigma_{p,d}^\Gamma(\epsilon, \chi) \in [0, 1]$ is the normalized area of the payoff region within class Γ for which p is stable under d – equivalently, the probability that p is stable under d when a payoff configuration is drawn uniformly at random from $\mathcal{G}_\Gamma$. For instance, $\sigma_{p,d}^{\text{SG}} = 1$ indicates that p is stable under d for every snowdrift game in the admissible parameter region, whereas $\sigma_{p,d}^{\text{PD}} = 1/2$ indicates stability on exactly one half of the normalized prisoner’s dilemma payoff space.

To facilitate comparison of moral systems across the three classes of cooperative dilemmas, we also report the average stability index

$$\sigma_{p,d} = \frac{1}{3}\sigma_{p,d}^{\text{PD}} + \frac{1}{3}\sigma_{p,d}^{\text{SG}} + \frac{1}{3}\sigma_{p,d}^{\text{SH}}. \quad (32)$$

This quantity assigns equal weight to each of the three dilemma classes, regardless of the differing areas of their payoff regions.

Below, we report $\sigma_{p,d}^\Gamma$ for each $\Gamma \in \{\text{PD}, \text{SG}, \text{SH}\}$ and the average stability index (32) for certain moral systems (p, d) in the absence of execution errors ($\epsilon = 0$) and in the limit of vanishing assignment errors ($\chi \rightarrow 0$). Writing $A_{\Gamma,p,d} := \lambda_2(\mathcal{S}_{p,d}^\Gamma)$ and $A_\Gamma := \lambda_2(\mathcal{G}_\Gamma)$ for the corresponding areas, these stability indices can be expressed as

$$\lim_{\chi \rightarrow 0} (\sigma_{p,d}^\Gamma)_{\epsilon=0} = \frac{\lim_{\chi \rightarrow 0} (A_{\Gamma,p,d})_{\epsilon=0}}{A_\Gamma}. \quad (33)$$

In the cases we study, the stability regions $\mathcal{S}_{p,d}^\Gamma$ are polygons whose vertices can be determined analytically; their areas are then computed with the shoelace formula:

$$A = \frac{1}{2} \sum_{i=0}^{n-1} (x_i y_{i+1} - x_{i+1} y_i), \quad (34)$$

where $(x_0, y_0), (x_1, y_1), \dots, (x_n, y_n)$ are the vertices listed in counter-clockwise order, with $x_n = x_0$ and $y_n = y_0$.

This polygonal structure of $\mathcal{S}_{p,d}^\Gamma$, on which the analytical stability indices reported below rely, holds in the absence of execution errors and in the limit of vanishing assignment errors ($\epsilon = 0, \chi \rightarrow 0$). At the finite error rates used for our numerical results ($\epsilon = \chi = 0.02$ in the main text, Fig. 2; and the error-rate grid of Section 4), the stability regions $\mathcal{S}_{p,d}^\Gamma$ are generally no longer polygons. There, we estimate each stability index by Monte Carlo integration, as detailed in Section 4.2.

1.10.1 Areas of the payoff regions

We compute the total area $A_\Gamma = \lambda_2(\mathcal{G}_\Gamma)$ for each game class. In each case, $\mathcal{G}_\Gamma$ is a convex polygon determined by the normalization constraints (Eqs. (13)–(15)).

Prisoner’s dilemmas. The two free parameters are R and P , with constraints given by (13). The region $\mathcal{G}_{PD}$ (Extended Data Fig. 2a) is a quadrilateral whose vertices, in (R, P) coordinates, are

$$\left(\frac{1}{2}, 0\right), (1, 0), (1, 1), \left(\frac{1}{2}, \frac{1}{2}\right).$$

Applying the shoelace formula (34) gives

$$A_{PD} = \frac{3}{8}. \quad (35)$$

Snowdrift games. The two free parameters are R and S , with constraints given by (14). The region $\mathcal{G}_{SG}$ (Extended Data Fig. 2b) is a triangle whose vertices, in (R, S) coordinates, are

$$\left(\frac{1}{2}, 0\right), (1, 0), (1, 1).$$

Applying the shoelace formula (34) gives

$$A_{SG} = \frac{1}{4}. \quad (36)$$

Stag hunts. The two free parameters are T and P , with constraints given by (15). The region $\mathcal{G}_{SH}$ (Extended Data Fig. 2c) is a triangle whose vertices, in (T, P) coordinates, are

$$(0, 0), (1, 0), (1, 1).$$

Applying the shoelace formula (34) gives

$$A_{SH} = \frac{1}{2}. \quad (37)$$

1.10.2 Stability indices for two recurring binding conditions

Only two non-trivial stability regions arise in the stability analyses that follow (Sections 3.6 and 3.7); we compute their stability indices here.

Snowdrift games with $R + P > 2S$. Consider a snowdrift game (10) and a moral system (p, d) such that p is stable under d for $\epsilon = 0$ and $\chi \rightarrow 0$ if and only if $R + P > 2S$ holds. Under the normalization (14), this condition becomes $R > 2S$, so the stability region is

$$\mathcal{S}_{p,d}^{\text{SG}} = \{ \pi \in \mathcal{G}_{\text{SG}} : R > 2S \},$$

a quadrilateral (Fig. 1a) whose vertices, in (R, S) coordinates, are

$$\left(\frac{1}{2}, 0\right), (1, 0), \left(1, \frac{1}{2}\right), \left(\frac{2}{3}, \frac{1}{3}\right).$$

Applying the shoelace formula (34) gives

$$\lim_{\chi \rightarrow 0} (A_{\text{SG}, p, d})_{\epsilon=0} = \frac{1}{6}. \quad (38)$$

Substituting (38) and (36) into (33) yields the following lemma.

Lemma 1 (Stability index across snowdrift games with binding condition $R + P > 2S$). *Let $\epsilon = 0$ and $\chi \rightarrow 0$. If p is stable under d across snowdrift games if and only if $R + P > 2S$, then $\lim_{\chi \rightarrow 0} (\sigma_{p,d}^{\text{SG}})_{\epsilon=0} = 2/3$.*

Prisoner's dilemmas with $2R > T + P$. Consider a prisoner's dilemma (9) and a moral system (p, d) such that p is stable under d for $\epsilon = 0$ and $\chi \rightarrow 0$ if and only if $2R > T + P$ holds. Under the normalization (13), this condition becomes $P < 2R - 1$, so the stability region is

$$\mathcal{S}_{p,d}^{\text{PD}} = \{ \pi \in \mathcal{G}_{\text{PD}} : P < 2R - 1 \},$$

a triangle (Fig. 1b) whose vertices, in (R, P) coordinates, are

$$\left(\frac{1}{2}, 0\right), (1, 0), (1, 1).$$

Applying the shoelace formula (34) gives

$$\lim_{\chi \rightarrow 0} (A_{\text{PD}, p, d})_{\epsilon=0} = \frac{1}{4}. \quad (39)$$

Substituting (39) and (35) into (33) yields the following lemma.

Lemma 2 (Stability index across prisoner's dilemmas when $2R > T + P$ binds). *Let $\epsilon = 0$ and $\chi \rightarrow 0$. If p is stable under d across prisoner's dilemmas if and only if $2R > T + P$, then $\lim_{\chi \rightarrow 0} (\sigma_{p,d}^{\text{PD}})_{\epsilon=0} = 2/3$.*

1.10.3 Weighted average stability indices

More generally, we can define a weighted average stability index

$$\sigma_{p,d}^{\text{w}} = w_{\text{PD}} \sigma_{p,d}^{\text{PD}} + w_{\text{SG}} \sigma_{p,d}^{\text{SG}} + w_{\text{SH}} \sigma_{p,d}^{\text{SH}}$$

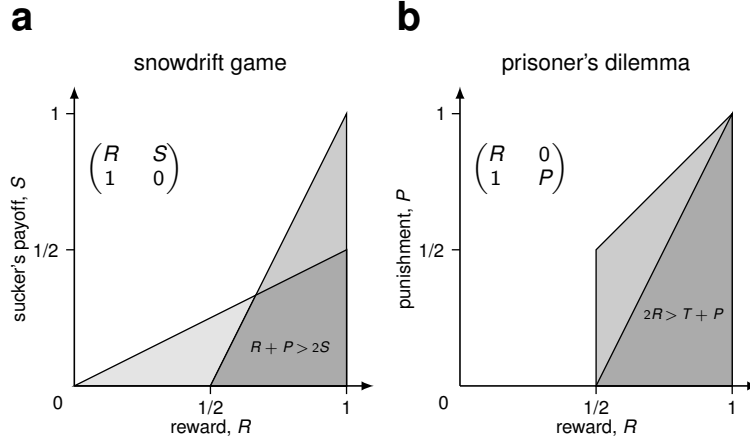


Figure 1: Illustration of Lemmas 1 and 2. **a**, The region where $R + P > 2S$ is superimposed on the region of snowdrift games, in normalized payoff space (Lemma 1). **b**, The region where $2R > T + P$ is superimposed on the region of prisoner's dilemmas in normalized payoff space (Lemma 2).

for weights $w_{PD}, w_{SG}, w_{SH} > 0$ with $w_{PD} + w_{SG} + w_{SH} = 1$. The average stability index (32) is the particular case of equal weights, $w_{PD} = w_{SG} = w_{SH} = 1/3$. An alternative weights each class in proportion to the area of its payoff region,

$$w_{\Gamma} = \frac{A_{\Gamma}}{\sum_{\Gamma} A_{\Gamma}} \text{ for } \Gamma \in \{PD, SG, SH\},$$

which, using the areas $A_{PD} = 3/8$, $A_{SG} = 1/4$, $A_{SH} = 1/2$ computed above, gives

$$w_{PD} = \frac{1}{3}, \quad w_{SG} = \frac{2}{9}, \quad \text{and} \quad w_{SH} = \frac{4}{9}.$$

The prisoner's dilemma thus receives the same weight, $1/3$, under both the equal-weight and area-weighted schemes; the two differ only in how they split the remaining weight between snowdrift games and stag hunts. We adopt this area-weighted index as a robustness check on our main results (Figs. 26 and 27).

2 Definitions and preliminary results

Here we introduce terminology for the subsets of strategies and moral systems that recur in our analysis. These definitions provide the conceptual scaffolding for the theoretical results of subsequent sections.

In Section 2.1, we distinguish strategies according to whether they respond to reputation information. Unconditional strategies (ALLC and ALLD) prescribe the same action regardless of the two players' reputations, whereas conditional strategies allow behavior to depend on this information. Among conditional strategies, we single out the discriminating strategies, which cooperate with co-players regarded as good and defect against those regarded as bad (provided the focal player is in good standing), and the contrite strategies, which cooperate whenever the focal player is in bad standing. We extend these notions to moral systems (p, d) , defining conditional, discriminating, and contrite systems accordingly.

In Section 2.2, we characterize moral systems by their equilibrium properties at the population level. A moral system is self-approving if a resident population comes to regard itself as good,

and self-cooperative if it sustains mutual cooperation, both in the absence of execution errors and the limit of vanishing assignment errors. Combining these, a system is righteous if it is both self-approving and self-cooperative. Finally, a system is universally stable if it resists invasion by any alternative strategy across all three classes of cooperative dilemma.

In Section 2.3, we introduce consistent systems, in which the assessment rule endorses as good exactly those who follow the action rule and penalizes those who deviate from it. Consistent discriminating systems are central to our analysis; we describe their four subsets, one for each discriminating strategy.

In Section 2.4, we define the leading-eight moral systems of Ohtsuki and Iwasa², which serve as reference points throughout our analysis. Finally, in Section 2.5, we formalize the mirror operation – the swapping of the two reputation labels – and prove the preliminary results used in later sections: a moral system and its mirror are behaviorally equivalent, whereas label-dependent properties such as self-approval hold for at most one member of each mirror pair.

2.1 Conditional, discriminating, and contrite strategies and systems

We begin by defining unconditional and conditional strategies.

Definition 1 (Unconditional strategies). *A strategy is said to be unconditional if $p_{BB} = p_{BG} = p_{GB} = p_{GG}$ holds.*

There are exactly two unconditional strategies in our strategy space: the ‘unconditional defector’ (ALLD), with $p_{BB} = p_{BG} = p_{GB} = p_{GG} = D$, and the ‘unconditional cooperator’ (ALLC), with $p_{BB} = p_{BG} = p_{GB} = p_{GG} = C$.

Definition 2 (Conditional strategies). *A strategy is said to be conditional if it is not unconditional.*

Among conditional strategies, a particularly important subset consists of those that, when the focal player is in good standing ($r_1 = G$), prescribe cooperation toward co-players regarded as good ($r_2 = G$) and defection against those regarded as bad ($r_2 = B$). We define these strategies formally as follows.

Definition 3 (Discriminating strategies). *A strategy p is said to be discriminating if both*

$$p_{GB} = D \text{ and } p_{GG} = C$$

hold.

As a string, discriminating strategies take the form $p = **DC$. Depending on what they prescribe when the focal player is in bad standing, we distinguish four discriminating strategies:

1. ‘(Standard) discriminator’ (DISC, $p = DCDC$), which, when in bad standing, discriminates by the co-player’s reputation just as in good standing.
2. ‘Contrite discriminator’ (CDISC, $p = CCDC$), which, when in bad standing, cooperates regardless of the co-player’s reputation.
3. ‘Obfuscated discriminator’ (ODISC, $p = CDDC$), which, when in bad standing, reverses the discriminating logic, cooperating with bad co-players and defecting against good ones.
4. ‘Petulant discriminator’ (PDISC, $p = DDDC$), which, when in bad standing, defects regardless of the co-player’s reputation.

ID	Bitstring	Name	Conditional?	Discriminating?	Contrite?
0	DDDD	ALLD	✗	✗	✗
1	CDDD		✓	✗	✗
2	DCDD		✓	✗	✗
3	CCDD		✓	✗	✓
4	DDCD		✓	✗	✗
5	CDCD		✓	✗	✗
6	DCCD		✓	✗	✗
7	CCCD	PDISC	✓	✗	✓
8	DDDC		✓	✓	✗
9	CDDC		✓	✓	✗
10	DCDC		✓	✓	✗
11	CCDC	DISC	✓	✓	✓
12	DDCC		✓	✗	✗
13	CDCC	CDISC	✓	✗	✗
14	DCCC		✓	✗	✗
15	CCCC	ALLC	✗	✗	✓

Table 1: The 16 strategies (action rules). Each strategy is a Boolean function written as a bitstring $p = p_{BB}p_{BG}p_{GB}p_{GG}$ in the order BB, BG, GB, GG; its ID is this bitstring read as a little-endian binary number (C = G = 1, D = B = 0, with BB the least significant bit). The last three columns indicate whether the strategy is conditional, discriminating, and contrite (✓) or not (✗).

A further property central to our analysis concerns strategies that prescribe cooperation whenever the focal player is in bad standing ($r_1 = B$), a readiness to make amends that we call contrition.

Definition 4 (Contrite strategies). *A strategy p is said to be contrite if both*

$$p_{BB} = C \text{ and } p_{BG} = C$$

hold.

As a string, contrite strategies take the form $p = CC^{**}$. Among the four discriminating strategies, CDISC ($p = CCDC$) is the only one that is also contrite; it is thus the unique contrite discriminator.

All $2^4 = 16$ strategies, along with their conditional, discriminating, and contrite status, are shown in Table 1.

Our general goal is to identify moral systems (p, d) that sustain high levels of cooperation (as measured by the cooperation index) and resist invasion by every alternative strategy $\tilde{p} \neq p$. Neither unconditional strategy yields such a system, whatever the assessment rule d : ALLD ($p = DDDD$) sustains no cooperation, since individuals following it never cooperate, while ALLC ($p = CCCC$) is invaded by ALLD in prisoner's dilemmas and snowdrift games¹. For these reasons, we focus on moral systems implementing a conditional strategy.

Definition 5 (Conditional systems). *A moral system (p, d) is said to be conditional if its action rule p is conditional.*

We denote the set of conditional systems by C . Explicitly,

$$C = \{(p, d) \mid p \notin \{\text{ALLC}, \text{ALLD}\}\}.$$

Since conditional systems implementing a discriminating strategy play a central role in our analysis, we single them out with the following definition.

Definition 6 (Discriminating systems). *A moral system (p, d) is said to be discriminating if its action rule p is discriminating.*

We denote the set of discriminating systems by $\mathcal{D}$. Explicitly,

$$\mathcal{D} = \{(p, d) \mid p \in \{\text{DISC}, \text{CDISC}, \text{ODISC}, \text{PDISC}\}\}.$$

Analogously, we single out systems whose action rule is contrite.

Definition 7 (Contrite systems). *A moral system (p, d) is said to be contrite if its action rule p is contrite.*

Since CDISC is the unique contrite discriminating strategy, a discriminating system is contrite if and only if its action rule is CDISC.

2.2 Self-improving, self-cooperative, righteous, and universally stable systems

We begin by defining properties of moral systems related to their reputation equilibria.

Definition 8 (Self-improving systems). *A conditional system (p, d) is said to be self-improving if its resident reputation equilibrium in the absence of execution errors approaches one in the limit of vanishing assignment errors, that is, if*

$$\lim_{\chi \rightarrow 0} \left(g_{p,d}^* \right)_{\epsilon=0} = 1 \quad (40)$$

holds. If (p, d) is self-improving, we also say that p is self-improving under d .

We denote the set of self-improving systems by $\mathcal{A}$. Thus,

$$\mathcal{A} = \left\{ (p, d) \mid (p, d) \in C \text{ and } \lim_{\chi \rightarrow 0} \left(g_{p,d}^* \right)_{\epsilon=0} = 1 \right\}.$$

We also define the converse property, in which a resident population comes to regard itself as bad at equilibrium.

Definition 9 (Self-disimproving systems). *A conditional system (p, d) is said to be self-disimproving if its resident reputation equilibrium in the absence of execution errors approaches zero in the limit of vanishing assignment errors, that is, if*

$$\lim_{\chi \rightarrow 0} \left(g_{p,d}^* \right)_{\epsilon=0} = 0 \quad (41)$$

holds. If (p, d) is self-disimproving, we also say that p is self-disimproving under d .

We denote the set of self-disimproving systems by $\mathcal{Z}$. Thus,

$$\mathcal{Z} = \left\{ (p, d) \mid (p, d) \in C \text{ and } \lim_{\chi \rightarrow 0} \left(g_{p,d}^* \right)_{\epsilon=0} = 0 \right\}.$$

We now define an analogous property concerning cooperation levels rather than reputation equilibria, using the cooperation index (Section 1.7), which measures how frequently individuals cooperate.

Definition 10 (Self-cooperative systems). A conditional system (p, d) is said to be self-cooperative if its cooperation index in the absence of execution errors, denoted by $(\kappa_{p,d})_{\epsilon=0}$, approaches one in the limit of vanishing assignment errors, that is, if

$$\lim_{\chi \rightarrow 0} (\kappa_{p,d})_{\epsilon=0} = 1$$

holds. If (p, d) is self-cooperative, we also say that p is self-cooperative under d .

We denote the set of self-cooperative systems by $\mathcal{K}$. Thus,

$$\mathcal{K} = \left\{ (p, d) \mid (p, d) \in \mathcal{C} \text{ and } \lim_{\chi \rightarrow 0} (\kappa_{p,d})_{\epsilon=0} = 1 \right\}.$$

Combining self-approval and self-cooperation yields the notion of a righteous system.

Definition 11 (Righteous systems). A conditional system (p, d) is said to be righteous if it is both self-approving and self-cooperative. If (p, d) is righteous, we also say that p is righteous under d .

We denote the set of righteous systems by $\mathcal{R}$. Thus,

$$\mathcal{R} = \mathcal{A} \cap \mathcal{K}. \quad (42)$$

A population implementing a righteous system (p, d) is sensitive to reputation information (since p is conditional), mostly regards itself as good (since (p, d) is self-approving), and mostly cooperates (since (p, d) is self-cooperative).

The definitions above characterize a moral system's equilibrium behavior in a resident population. We now turn to its robustness against invasion by alternative strategies.

Definition 12 (Universally stable systems). A moral system (p, d) is said to be universally stable if its stability indices in the absence of execution errors, denoted by $(\sigma_{p,d}^{\text{PD}})_{\epsilon=0}$, $(\sigma_{p,d}^{\text{SG}})_{\epsilon=0}$, and $(\sigma_{p,d}^{\text{SH}})_{\epsilon=0}$, all approach one in the limit of vanishing assignment errors, that is, if

$$\lim_{\chi \rightarrow 0} (\sigma_{p,d}^{\text{PD}})_{\epsilon=0} = \lim_{\chi \rightarrow 0} (\sigma_{p,d}^{\text{SG}})_{\epsilon=0} = \lim_{\chi \rightarrow 0} (\sigma_{p,d}^{\text{SH}})_{\epsilon=0} = 1$$

holds. If a moral system (p, d) is universally stable, we also say that p is universally stable under d .

By Definition 12, if (p, d) is universally stable, then p is stable under d across all cooperative dilemmas, in the sense of resisting invasion by any alternative strategy $\tilde{p} \neq p$ for any payoff matrix we consider, in the absence of execution errors ($\epsilon = 0$) and in the limit of vanishing assignment errors ($\chi \rightarrow 0$). If a system (p, d) is universally stable, then its weighted average stability index satisfies

$$\lim_{\chi \rightarrow 0} \sigma_{p,d}^{\mathbf{w}} = 1$$

for any vector of weights $(\mathbf{w}_{\text{PD}}, \mathbf{w}_{\text{SG}}, \mathbf{w}_{\text{SH}})$.

2.3 Consistent systems and consistent discriminating systems

2.3.1 Consistent systems

Consistency captures the idea that a moral system's action rule and assessment rule do not contradict each other: following the prescribed behavior is rewarded with a good reputation, and deviating from it is punished with a bad one.

Definition 13 (Consistent systems). *A conditional system (p, d) is said to be consistent if both*

$$d_{BB} = d_{BG} = d_{GB} = d_{GG} = G, \text{ and} \quad (43a)$$

$$\bar{d}_{BB} = \bar{d}_{BG} = \bar{d}_{GB} = \bar{d}_{GG} = B \quad (43b)$$

hold. If (p, d) is consistent, we also say that p is consistent with d , or that d is consistent with p .

We denote the set of consistent systems by $\mathcal{Q}$. Thus,

$$\mathcal{Q} = \{(p, d) \mid (p, d) \in \mathcal{C}, d_{r_1 r_2} = G, \text{ and } \bar{d}_{r_1 r_2} = B \text{ for all } r_1, r_2 \in \{B, G\}\}.$$

Recall our notational conventions for $d_{r_1 r_2}$ (16d) and $\bar{d}_{r_1 r_2}$ (16f). Conditions (43a) and (43b) together say that, whenever the co-player plays according to action rule p , assessment rule d regards the focal player as good if they follow p 's prescription and as bad if they deviate from it. The moral system (p, d) is then consistent in the sense that there is no contradiction between the behavior prescribed by the action rule and the moral judgment expressed by the assessment rule.

For each strategy p , the two conditions in Eq. (43) specify the values of eight out of the sixteen bits of an assessment rule that is consistent with p . Once these values are set, the remaining eight bits can be set to any binary value, and d will still be consistent with p , according to Definition 13. It follows that each strategy p gives rise to $2^8 = 256$ consistent systems (p, d) .

Remark 1 (Equivalence of consistent systems at $\epsilon = 0$). *The eight unconstrained bits of a consistent assessment rule are precisely those of the form $d(r_1, r_2, a_1, \neg p(r_2, r_1))$ for $r_1, r_2 \in \{B, G\}$ and $a_1 \in \{D, C\}$ – that is, the bits specifying how to assess a focal player when the co-player's action deviates from the resident prescription. In the reputation transition probability (19), these bits appear only in the terms proportional to $\epsilon(1 - \epsilon)$ and ϵ^2 , which vanish when $\epsilon = 0$. The same holds for the mutant reputation transition (24): at $\epsilon = 0$, the surviving term $d(r_1, r_2, \tilde{p}(r_1, r_2), p(r_2, r_1))$ involves a constrained bit regardless of whether the mutant's action agrees with the resident's, since conditions (43a) and (43b) together fix $d(r_1, r_2, a_1, p(r_2, r_1))$ for both values of a_1 . Consequently, all 256 consistent systems sharing the same action rule p have identical resident equilibria $g_{p,d}^*$, mutant equilibria $h_{\tilde{p},p,d}^*$, payoffs, and hence stability and cooperation indices, whenever $\epsilon = 0$.*

2.3.2 Consistent discriminating systems

Of special interest are consistent systems whose action rule is also discriminating.

Definition 14 (Consistent discriminating systems). *A conditional system (p, d) is said to be consistent discriminating if it is both discriminating and consistent.*

Thus,

$$\mathcal{Q}_{\mathcal{D}} = \mathcal{Q} \cap \mathcal{D}.$$

Depending on the action rule, $\mathcal{Q}_{\mathcal{D}}$ can be partitioned into four subsets:

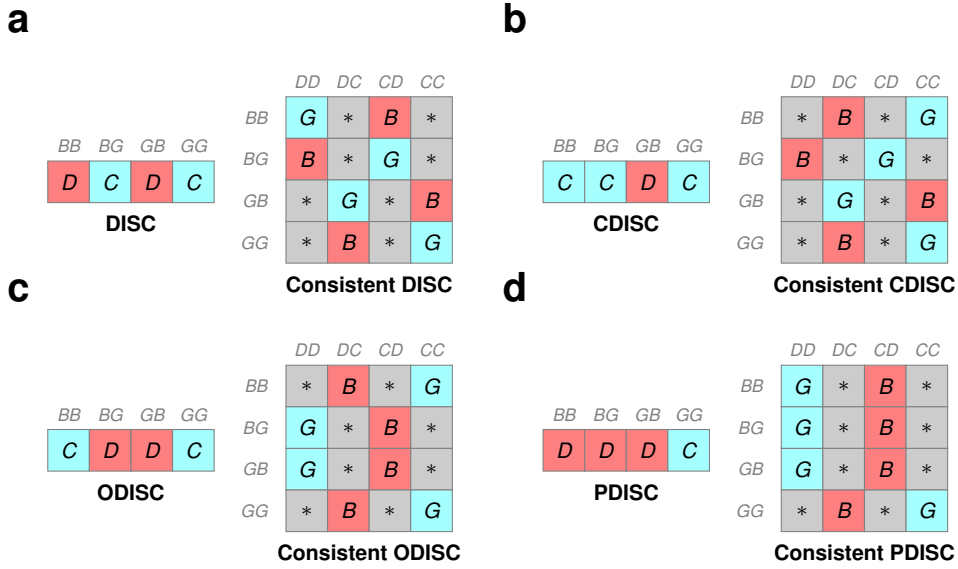


Figure 2: Action and assessment rules of the four subsets of consistent discriminating systems.

1. ‘Standard consistent discriminating systems’ ($p = \text{DCDC}$, subset $\mathcal{Q}_{\text{DISC}}$):

$$\mathcal{Q}_{\text{DISC}} = \{(p, d) \mid p = \text{DCDC}, d_{\text{BBCD}} = d_{\text{BGDD}} = d_{\text{GBCC}} = d_{\text{GGDC}} = B, \text{ and } d_{\text{BBDD}} = d_{\text{BGCD}} = d_{\text{GBDC}} = d_{\text{GGCC}} = G\}. \quad (44)$$

2. ‘Contrite consistent discriminating systems’ ($p = \text{CCDC}$, subset $\mathcal{Q}_{\text{CDISC}}$):

$$\mathcal{Q}_{\text{CDISC}} = \{(p, d) \mid p = \text{CCDC}, d_{\text{BBDC}} = d_{\text{BGDD}} = d_{\text{GBCC}} = d_{\text{GGDC}} = B, \text{ and } d_{\text{BBCC}} = d_{\text{BGCD}} = d_{\text{GBDC}} = d_{\text{GGCC}} = G\}. \quad (45)$$

3. ‘Obfuscated consistent discriminating systems’ ($p = \text{CDDC}$, subset $\mathcal{Q}_{\text{ODISC}}$):

$$\mathcal{Q}_{\text{ODISC}} = \{(p, d) \mid p = \text{CDDC}, d_{\text{BBDC}} = d_{\text{BGCD}} = d_{\text{GBCD}} = d_{\text{GGDC}} = B, \text{ and } d_{\text{BBCC}} = d_{\text{BGDD}} = d_{\text{GBDD}} = d_{\text{GGCC}} = G\}. \quad (46)$$

4. ‘Petulant consistent discriminating systems’ ($p = \text{DDDC}$, subset $\mathcal{Q}_{\text{PDISC}}$):

$$\mathcal{Q}_{\text{PDISC}} = \{(p, d) \mid p = \text{DDDC}, d_{\text{BBCD}} = d_{\text{BGCD}} = d_{\text{GBCD}} = d_{\text{GGDC}} = B, \text{ and } d_{\text{BBDD}} = d_{\text{BGDD}} = d_{\text{GBDD}} = d_{\text{GGCC}} = G\}. \quad (47)$$

The action and assessment rules of each of these subsets of consistent discriminating systems are shown in Fig. 2.

2.4 The leading eight

The leading eight, first identified by Ohtsuki and Iwasa², are a set of eight moral systems. Each uses either DISC or CDISC as its action rule, together with a third-order assessment rule of type $a_1 r_1 r_2$ – one that ignores the co-player’s action a_2 – satisfying

$$d_{\text{BGD}*} = d_{\text{GGD}*} = B, \quad (48a)$$

$$d_{\text{BGC}*} = d_{\text{GGC}*} = G, \quad (48b)$$

$$d_{\text{GBD}*} = G. \quad (48c)$$

Group I

	DD	DC	CD	CC
BB	B	B	G	G
BG	B	B	G	G
GB	G	G	G	G
GG	B	B	G	G

L1 (Standing)

	DD	DC	CD	CC
BB	B	B	G	G
BG	B	B	G	G
GB	G	G	B	B
GG	B	B	G	G

L2 (Consistent Standing)

Group II

	DD	DC	CD	CC
BB	G	G	G	G
BG	B	B	G	G
GB	G	G	G	G
GG	B	B	G	G

L3 (Simple Standing)

	DD	DC	CD	CC
BB	G	G	B	B
BG	B	B	G	G
GB	G	G	G	G
GG	B	B	G	G

L4 (SS-SJ)

	DD	DC	CD	CC
BB	G	G	G	G
BG	B	B	G	G
GB	G	G	B	B
GG	B	B	G	G

L5 (SJ-SS)

	DD	DC	CD	CC
BB	G	G	B	B
BG	B	B	G	G
GB	G	G	B	B
GG	B	B	G	G

L6 (Stern Judging)

Group III

	DD	DC	CD	CC
BB	B	B	B	B
BG	B	B	G	G
GB	G	G	G	G
GG	B	B	G	G

L7 (Staying)

	DD	DC	CD	CC
BB	B	B	B	B
BG	B	B	G	G
GB	G	G	B	B
GG	B	B	G	G

L8 (Judging)

Figure 3: Social norms of the leading-eight moral systems, arranged by group. Group I systems (L1 and L2) use CDISC as their action rule; Group II (L3–L6) and Group III (L7 and L8) systems use DISC.

Conditions (48a) and (48b) state that, irrespective of the focal player’s reputation, cooperation with good co-players is judged as good and defection against them as bad. Condition (48c) states that defection against bad co-players by good individuals is judged as good, encapsulating the concept of ‘justified defection’². The leading-eight moral systems are named and described in Fig. 3.

The leading-eight moral systems differ in two respects: (i) whether they use DISC or CDISC as their action rule, and (ii) their assignment of the bits d_{BBD*} , d_{BBC*} , and d_{GBC*} , which govern how interactions involving bad co-players are judged. In line with the original analysis by Ohtsuki and Iwasa², we distinguish the following three groups.

Group I. Group I systems use $p = CCDC$ (i.e. CDISC), with $d_{BBD*} = B$ and $d_{BBC*} = G$. We denote this set by $\mathcal{L}_I$:

$$\mathcal{L}_I = \{(p, d) \mid p = CCDC, d_{BBD*} = d_{BGD*} = d_{GGD*} = B, \\ d_{BBC*} = d_{BGC*} = d_{GBD*} = d_{GGC*} = G, \text{ and } d_{GBCD} = d_{GBCC}\} . \quad (49)$$

The two members are L1 (CDISC, Standing), with $d_{GBC*} = G$, and L2 (CDISC, Consistent Standing), with $d_{GBC*} = B$ (see Fig. 3):

$$\begin{aligned} L1 = \{ (p, d) \mid p = \text{CCDC}, d_{BBD*} = d_{BGD*} = d_{GGD*} = B, \\ \text{and } d_{BBC*} = d_{BGC*} = d_{GBD*} = d_{GBC*} = d_{GGC*} = G \}, \end{aligned} \quad (50)$$

$$\begin{aligned} L2 = \{ (p, d) \mid p = \text{CCDC}, d_{BBD*} = d_{BGD*} = d_{GBC*} = d_{GGD*} = B, \\ \text{and } d_{BBC*} = d_{BGC*} = d_{GBD*} = d_{GGC*} = G \}. \end{aligned} \quad (51)$$

Note that L2 belongs to $\mathcal{Q}_{\text{CDISC}}$ (cf. Eqs. (51) and (45)).

Group II. Group II systems use $p = \text{DCDC}$ (i.e. DISC), with $d_{BBD*} = G$. We denote this set by $\mathcal{L}_{\text{II}}$:

$$\begin{aligned} \mathcal{L}_{\text{II}} = \{ (p, d) \mid p = \text{DCDC}, d_{BGD*} = d_{GGD*} = B, \\ d_{BBD*} = d_{BGC*} = d_{GBD*} = d_{GGC*} = G, \\ d_{BBCD} = d_{BBCC}, \text{ and } d_{GBCD} = d_{GBCC} \}. \end{aligned} \quad (52)$$

The four members are L3 (DISC, Simple Standing), with $d_{BBC*} = d_{GBC*} = G$; L4 (DISC, SS-SJ), with $d_{BBC*} = B$ and $d_{GBC*} = G$; L5 (DISC, SJ-SS), with $d_{BBC*} = G$ and $d_{GBC*} = B$; and L6 (DISC, Stern Judging), with $d_{BBC*} = d_{GBC*} = B$ (see Fig. 3):

$$\begin{aligned} L3 = \{ (p, d) \mid p = \text{DCDC}, d_{BGD*} = d_{GGD*} = B, \\ \text{and } d_{BBD*} = d_{BBC*} = d_{BGC*} = d_{GBD*} = d_{GBC*} = d_{GGC*} = G \}, \end{aligned} \quad (53)$$

$$\begin{aligned} L4 = \{ (p, d) \mid p = \text{DCDC}, d_{BBC*} = d_{BGD*} = d_{GGD*} = B, \\ \text{and } d_{BBD*} = d_{BGC*} = d_{GBD*} = d_{GBC*} = d_{GGC*} = G \}, \end{aligned} \quad (54)$$

$$\begin{aligned} L5 = \{ (p, d) \mid p = \text{DCDC}, d_{BGD*} = d_{GBC*} = d_{GGD*} = B, \\ \text{and } d_{BBC*} = d_{BBD*} = d_{BGC*} = d_{GBD*} = d_{GGC*} = G \}, \end{aligned} \quad (55)$$

$$\begin{aligned} L6 = \{ (p, d) \mid p = \text{DCDC}, d_{BBC*} = d_{BGD*} = d_{GBC*} = d_{GGD*} = B, \\ \text{and } d_{BBD*} = d_{BGC*} = d_{GBD*} = d_{GGC*} = G \}. \end{aligned} \quad (56)$$

Note that L6 belongs to $\mathcal{Q}_{\text{DISC}}$ (cf. Eqs. (56) and (44)).

Group III. Group III systems use $p = \text{DCDC}$ (i.e. DISC), with $d_{BBD*} = d_{BBC*} = B$. We denote this set by $\mathcal{L}_{\text{III}}$:

$$\begin{aligned} \mathcal{L}_{\text{III}} = \{ (p, d) \mid p = \text{DCDC}, d_{BBD*} = d_{BBC*} = d_{BGD*} = d_{GGD*} = B, \\ d_{BGC*} = d_{GBD*} = d_{GGC*} = G, \text{ and } d_{GBCD} = d_{GBCC} \}. \end{aligned} \quad (57)$$

The two members are L7 (DISC, Staying), with $d_{GBC*} = G$, and L8 (DISC, Judging), with $d_{GBC*} = B$ (see Fig. 3):

$$\begin{aligned} L7 = \{ (p, d) \mid p = \text{DCDC}, d_{BBD*} = d_{BBC*} = d_{BGD*} = d_{GGD*} = B, \\ \text{and } d_{BGC*} = d_{GBD*} = d_{GBC*} = d_{GGC*} = G \}, \end{aligned} \quad (58)$$

$$\begin{aligned} L8 = \{ (p, d) \mid p = \text{DCDC}, d_{BBD*} = d_{BBC*} = d_{BGD*} = d_{GBC*} = d_{GGD*} = B, \\ \text{and } d_{BGC*} = d_{GBD*} = d_{GGC*} = G \}. \end{aligned} \quad (59)$$

We denote by $\mathcal{L}$ the set of all leading-eight moral systems, so that

$$\mathcal{L} = \mathcal{L}_{\text{I}} \cup \mathcal{L}_{\text{II}} \cup \mathcal{L}_{\text{III}}. \quad (60)$$

2.5 Mirror systems

The reputation labels G (‘good’) and B (‘bad’) are arbitrary bookkeeping devices: unlike actions, which determine payoffs through the payoff function v , reputations have no intrinsic consequences. Swapping the two labels throughout a moral system therefore yields a system that is behaviorally indistinguishable from the original – an observation made by Ohtsuki and Iwasa² for the third-order assessment rules they studied, and which they termed ‘mirror symmetry’ (their Appendix A). This subsection formalizes this relabeling operation and verifies that the quantities of interest (reputation equilibria, cooperation index, stability indices) are invariant under it; Section 4.1 builds on these results to count moral systems up to relabeling and to specify the representatives evaluated in our numerical explorations.

Definition 15 (Mirror systems). *The mirror of a moral system (p, d) is the moral system $(\hat{p}, \hat{d})$ given by*

$$\hat{p}(r_1, r_2) = p(\neg r_1, \neg r_2), \quad (61a)$$

$$\hat{d}(r_1, r_2, a_1, a_2) = \neg d(\neg r_1, \neg r_2, a_1, a_2), \quad (61b)$$

for all $r_1, r_2 \in \{B, G\}$ and $a_1, a_2 \in \{D, C\}$, where $\neg$ denotes the swap $B \leftrightarrow G$ (equivalently, with our labeling convention (1), $\neg x = 1 - x$). A moral system that is equal to its mirror is said to be a self-mirror.

The mirror relabels reputations wherever they occur – in the reputation arguments of p and d , and in the reputation prescribed by d – while leaving actions untouched. As a bitstring (Table 1), $\hat{p}$ is simply p read from right to left, so that, for instance, the mirror of DISC ($p = \text{DCDC}$) is the strategy CDCD, which cooperates with co-players in bad standing. Applying the mirror twice recovers the original system, so the mirror map is an involution on the space of moral systems: it partitions this space into ‘mirror classes’ of size one (self-mirrors) or two (‘mirror pairs’).

The next lemma shows that the mirror system reproduces the reputation dynamics of the original system with the roles of G and B interchanged.

Lemma 3 (Reputation dynamics of mirror systems). *Let $(\hat{p}, \hat{d})$ be the mirror of (p, d) , and let $\hat{p}$ be the mirror of a mutant strategy $\tilde{p}$. Denote by $\hat{\delta}_{r_1 r_2}$ the probability (19) that a resident of a population implementing $(\hat{p}, \hat{d})$ is assigned a good reputation, and by $\hat{\delta}_{r_1 r_2}$ the corresponding probability (24) for a mutant adopting $\hat{p}$ in this population. Then, for all error rates $\chi \in (0, 1/2)$ and $\epsilon \in [0, 1/2)$, and for all $r_1, r_2 \in \{B, G\}$,*

$$\hat{\delta}_{r_1 r_2} = 1 - \delta_{\neg r_1 \neg r_2} \quad \text{and} \quad \hat{\delta}_{r_1 r_2} = 1 - \tilde{\delta}_{\neg r_1 \neg r_2}. \quad (62)$$

Consequently, the resident and mutant reputation equilibria satisfy

$$g_{\hat{p}, \hat{d}}^* = 1 - g_{p, d}^* \quad \text{and} \quad h_{\hat{p}, \hat{p}, \hat{d}}^* = 1 - h_{\tilde{p}, p, d}^*. \quad (63)$$

Proof. Substituting (61) into (19), each assessment value appearing in $\hat{\delta}_{r_1 r_2}$ equals one minus the corresponding value appearing in $\delta_{\neg r_1 \neg r_2}$. For instance, in the absence of execution errors, $\hat{d}(r_1, r_2, \hat{p}(r_1, r_2), \hat{p}(r_2, r_1)) = 1 - d(\neg r_1, \neg r_2, p(\neg r_1, \neg r_2), p(\neg r_2, \neg r_1))$; the same identity holds for the three remaining assessment values, in which execution errors replace one or both actions with D, since the mirror leaves actions – and hence these erroneous defections – unchanged. As the weights $(1 - \epsilon)^2$, $\epsilon(1 - \epsilon)$, $\epsilon(1 - \epsilon)$, and ϵ^2 sum to one, the term in braces in (19) for the mirror system equals one minus its counterpart for the original system evaluated at $(\neg r_1, \neg r_2)$,

whence $\hat{\delta}_{r_1 r_2} = \chi + (1 - 2\chi) \left(1 - \frac{\delta_{\neg r_1 \neg r_2} - \chi}{1 - 2\chi}\right) = 1 - \delta_{\neg r_1 \neg r_2}$. The same substitution in (24) yields the mutant counterpart.

For the resident equilibrium, write the right-hand side of the reputation dynamics (20) as $F_{p,d}(g)$; expanding, $F_{p,d}(g) = ag^2 + bg + c$, with coefficients given in Eq. (22). Using (62),

$$\begin{aligned} F_{\hat{p},\hat{d}}(1-g) &= (1 - \delta_{\text{BB}})(1-g)^2 + (2 - \delta_{\text{GB}} - \delta_{\text{BG}})g(1-g) + (1 - \delta_{\text{GG}})g^2 - (1-g) \\ &= 1 - [\delta_{\text{BB}}(1-g)^2 + (\delta_{\text{BG}} + \delta_{\text{GB}})g(1-g) + \delta_{\text{GG}}g^2] - 1 + g = -F_{p,d}(g). \end{aligned}$$

Hence, if $g(t)$ solves the reputation dynamics $\dot{g} = F_{p,d}(g)$ of (p, d) , then $\hat{g}(t) := 1 - g(t)$ satisfies $\dot{\hat{g}} = -\dot{g} = -F_{p,d}(g) = F_{\hat{p},\hat{d}}(\hat{g})$, that is, it solves the reputation dynamics of $(\hat{p}, \hat{d})$; the converse follows by applying the same argument to the mirror system, since the mirror map is an involution. The unique asymptotically stable equilibria (21) of the two systems thus satisfy $g_{\hat{p},\hat{d}}^* = 1 - g_{p,d}^*$. An identical computation applied to the mutant dynamics (25), with coefficients $\hat{\delta}_{r_1 r_2}$ and resident equilibrium $1 - g_{p,d}^*$, shows that $h(t)$ solves the mutant dynamics of $\tilde{p}$ in (p, d) if and only if $1 - h(t)$ solves those of $\hat{\tilde{p}}$ in $(\hat{p}, \hat{d})$, whence the second identity in (63). $\square$

With Lemma 3 in place, the behavioral equivalence of a moral system and its mirror follows readily.

Proposition 2 (Mirror invariance of payoffs, cooperation, and stability). *Let $(\hat{p}, \hat{d})$ be the mirror of (p, d) . Then, for all error rates $\chi \in (0, 1/2)$ and $\epsilon \in [0, 1/2)$ and all payoff configurations:*

- (i) $w_{\hat{p},\hat{d}} = w_{p,d}$, and $\tilde{w}_{\hat{p},\hat{p},\hat{d}} = \tilde{w}_{\tilde{p},p,d}$ for every mutant strategy $\tilde{p}$;
- (ii) $\kappa_{\hat{p},\hat{d}} = \kappa_{p,d}$;
- (iii) $(\hat{p}, \hat{d})$ is stable if and only if (p, d) is stable; consequently, $\sigma_{\hat{p},\hat{d}}^\Gamma = \sigma_{p,d}^\Gamma$ for each game class $\Gamma \in \{\text{PD}, \text{SG}, \text{SH}\}$, and the average stability indices likewise coincide.

Proof. The expected payoffs (27) and (29) involve reputations only through the actions prescribed by the strategies, so (61a) gives $\hat{\omega}_{r_1 r_2} = \omega_{\neg r_1 \neg r_2}$ and $\hat{\tilde{\omega}}_{r_1 r_2} = \tilde{\omega}_{\neg r_1 \neg r_2}$. Substituting these identities and (63) into (28), (30), and (23), each term with reputation profile (r_1, r_2) for the mirror system equals the term with profile $(\neg r_1, \neg r_2)$ for the original system (for instance, in (28), $(1-g)^2 \omega_{\text{BB}}$ evaluated at $g = 1 - g_{p,d}^*$ for the mirror system equals $(g_{p,d}^*)^2 \omega_{\text{GG}}$ for the original one), which proves (i) and (ii). For (iii), the mirror map is a bijection of the strategy space onto itself with $\tilde{p} \neq p$ if and only if $\hat{\tilde{p}} \neq \hat{p}$, so by (i) the stability condition (31) holds for (p, d) if and only if it holds for $(\hat{p}, \hat{d})$. Since this equivalence holds at every payoff configuration, the stability regions $\mathcal{S}_{p,d}^\Gamma$ and $\mathcal{S}_{\hat{p},\hat{d}}^\Gamma$ coincide, and so do the stability indices σ^Γ and their (weighted) averages. $\square$

Whereas quantities defined from behavior alone are mirror-invariant, properties whose definitions refer to the reputation labels are not. The following corollary records the consequences for the classes of moral systems defined above.

Corollary 1 (Mirrors and classes of moral systems). *Let $(\hat{p}, \hat{d})$ be the mirror of a conditional system (p, d) . Then:*

- (i) (p, d) is self-approving if and only if $(\hat{p}, \hat{d})$ is self-disapproving; in particular, the mirror map restricts to a bijection $\mathcal{A} \rightarrow \mathcal{Z}$, so $|\mathcal{A}| = |\mathcal{Z}|$.

- (ii) (p, d) is self-cooperative (resp. universally stable) if and only if $(\hat{p}, \hat{d})$ is self-cooperative (resp. universally stable).
- (iii) Every self-mirror satisfies $g_{p,d}^* = 1/2$ for all error rates $\chi \in (0, 1/2)$ and $\epsilon \in [0, 1/2)$; in particular, no self-mirror is self-approving or self-disapproving, and no righteous system is a self-mirror.

Proof. Claims (i) and (ii) follow from Lemma 3 and Proposition 2 by taking the limits in Definitions 8, 9, 10, and 12 (the mirror of a conditional system is conditional, since $\hat{p}$ is unconditional if and only if p is). For (iii), a self-mirror satisfies $g_{p,d}^* = g_{\hat{p},\hat{d}}^* = 1 - g_{p,d}^*$ by (63), hence $g_{p,d}^* = 1/2$; a righteous system is self-approving by Definition 11, so it cannot be a self-mirror. $\square$

3 Analytical results

This section derives the analytical results that underpin the main text's central claims. We proceed in two stages. The first stage characterizes the structural properties of moral systems. Sections 3.1–3.3 identify necessary and sufficient conditions for a system to be self-approving, self-cooperative, righteous, consistent, or consistent discriminating, and establish the inclusion relations among these sets. Section 3.4 then verifies that the leading-eight moral systems fall into the appropriate sets in this hierarchy. The second stage turns to evolutionary stability. Section 3.5 identifies which righteous systems are universally stable, establishing that universal stability is equivalent to being a contrite consistent discriminating system. Sections 3.6 and 3.7 then provide detailed stability analyses for all consistent discriminating systems (grouped by action rule) and for each of the leading-eight moral systems individually. Throughout, we focus on self-approving systems; self-disapproving systems are behaviorally equivalent under the mirror operation of Section 2.5 and require no separate treatment.

3.1 Characterization of self-approving systems

Here, we characterize the sets of self-approving and self-disapproving moral systems. We begin with the following result, which provides simple necessary and sufficient conditions for a moral system to be self-approving (or self-disapproving).

Proposition 3 (Characterization of self-approving and self-disapproving systems). *A conditional system (p, d) is self-approving if and only if*

- (i) $d_{GG} = G$, and
- (ii) $d_{BB} + d_{BG} + d_{GB} \geq 2$, i.e., at least two of $\{d_{BB}, d_{BG}, d_{GB}\}$ are G .

Similarly, a conditional system (p, d) is self-disapproving if and only if

- (i) $d_{BB} = B$, and
- (ii) $d_{BG} + d_{GB} + d_{GG} \leq 1$, i.e., at least two of $\{d_{BG}, d_{GB}, d_{GG}\}$ are B .

Proof. To identify the set of self-approving systems for which condition (40) in Definition 8 holds, and the set of self-disapproving systems for which condition (41) in Definition 9 holds, we set $\epsilon = 0$ in Eq. (19), so that Eq. (22) can be written as

$$\begin{aligned} a(\chi) &= (1 - 2\chi) [d_{BB} + d_{GG} - d_{BG} - d_{GB}], \\ b(\chi) &= (1 - 2\chi) [d_{BG} + d_{GB} - 2d_{BB}] - 1, \\ c(\chi) &= \chi + (1 - 2\chi)d_{BB}, \end{aligned}$$

where we have expressed a , b , and c as functions of χ , and where we have used the shorthand notation (16d). There are $2^4 = 16$ different possible values for the binary string $(d_{BB}, d_{BG}, d_{GB}, d_{GG})$, and so 16 different cases to analyze. We begin by noting that, if

$$d_{BB} + d_{GG} = d_{BG} + d_{GB} \quad (64)$$

holds, then $a(\chi) = 0$ holds for all χ . In this case,

$$\lim_{\chi \rightarrow 0} (g_{p,d}^*)_{\epsilon=0} = - \lim_{\chi \rightarrow 0} \frac{c(\chi)}{b(\chi)} = - \frac{c(0)}{b(0)} = \frac{d_{BB}}{1 + d_{BB} - d_{GG}}. \quad (65)$$

Depending on the values of d_{BB} and d_{GG} , we have four different cases.

1. If $(d_{BB}, d_{GG}) = (B, B)$, then (65) simplifies to 0.
2. If $(d_{BB}, d_{GG}) = (B, G)$, then (65) results in the indeterminate form $0/0$. Applying L'Hôpital's rule, we evaluate the limit as

$$\begin{aligned} \lim_{\chi \rightarrow 0} (g_{p,d}^*)_{\epsilon=0} &= - \lim_{\chi \rightarrow 0} \frac{c'(\chi)}{b'(\chi)} \\ &= - \frac{1 - 2d_{BB}}{-2(d_{BG} + d_{GB} - 2d_{BB})} = \frac{1 - 2d_{BB}}{2(d_{GG} - d_{BB})} = \frac{1}{2}. \end{aligned}$$

3. If $(d_{BB}, d_{GG}) = (G, B)$, then (65) simplifies to $1/2$.
4. If $(d_{BB}, d_{GG}) = (G, G)$, then (65) simplifies to 1.

Suppose now that (64) does not hold, that is, that

$$d_{BB} + d_{GG} \neq d_{BG} + d_{GB} \quad (66)$$

holds. In this case, $a(\chi) \neq 0$ holds for all $\chi \in (0, 1/2)$, and

$$\lim_{\chi \rightarrow 0} (g_{p,d}^*)_{\epsilon=0} = \lim_{\chi \rightarrow 0} \frac{-b(\chi) - \sqrt{b(\chi)^2 - 4a(\chi)c(\chi)}}{2a(\chi)}. \quad (67)$$

Here, we have two main cases.

1. If $d_{BB} = B$, then $c(0) = 0$, and Eq. (67) reduces to

$$\lim_{\chi \rightarrow 0} (g_{p,d}^*)_{\epsilon=0} = \lim_{\chi \rightarrow 0} \frac{-b(\chi) - \sqrt{b(\chi)^2}}{2a(\chi)}. \quad (68)$$

There are five different subcases for which both $d_{BB} = B$ and (66) hold.

- (a) If $(d_{BB}, d_{BG}, d_{GB}, d_{GG}) = (B, B, B, G)$, then $a(0) = 1$, $b(0) = -1$, and (68) simplifies to 0.
- (b) If $(d_{BB}, d_{BG}, d_{GB}, d_{GG}) = (B, B, G, B)$, then $a(0) = -1$, $b(0) = 0$, and (68) simplifies to 0.
- (c) If $(d_{BB}, d_{BG}, d_{GB}, d_{GG}) = (B, G, B, B)$, then $a(0) = -1$, $b(0) = 0$, and (68) simplifies to 0.
- (d) If $(d_{BB}, d_{BG}, d_{GB}, d_{GG}) = (B, G, G, B)$, then $a(0) = -2$, $b(0) = 1$, and (68) simplifies to $1/2$.
- (e) If $(d_{BB}, d_{BG}, d_{GB}, d_{GG}) = (B, G, G, G)$, then $a(0) = -1$, $b(0) = 1$, and (68) simplifies to 1.

2. If $d_{BB} = G$, then $c(0) = 1$, and (67) reduces to

$$\lim_{\chi \rightarrow 0} (g_{p,d}^*)_{\epsilon=0} = \lim_{\chi \rightarrow 0} \frac{-b(\chi) - \sqrt{b(\chi)^2 - 4a(\chi)}}{2a(\chi)}. \quad (69)$$

There are five different subcases for which both $d_{BB} = G$ and (66) hold.

- (a) If $(d_{BB}, d_{BG}, d_{GB}, d_{GG}) = (G, B, B, B)$, then $a(0) = 1$, $b(0) = -3$, and (69) simplifies to $\frac{3-\sqrt{5}}{2} \approx 0.382$.
- (b) If $(d_{BB}, d_{BG}, d_{GB}, d_{GG}) = (G, B, B, G)$, then $a(0) = 2$, $b(0) = -3$, and (69) simplifies to $\frac{1}{2}$.
- (c) If $(d_{BB}, d_{BG}, d_{GB}, d_{GG}) = (G, B, G, G)$, then $a(0) = 1$, $b(0) = -2$, and (69) simplifies to 1.
- (d) If $(d_{BB}, d_{BG}, d_{GB}, d_{GG}) = (G, G, B, G)$, then $a(0) = 1$, $b(0) = -2$, and (69) simplifies to 1.
- (e) If $(d_{BB}, d_{BG}, d_{GB}, d_{GG}) = (G, G, G, B)$, then $a(0) = b(0) = -1$, and (69) simplifies to $\frac{\sqrt{5}-1}{2} \approx 0.618$.

Table 2 shows a summary of the case distinctions and the results. By Definitions 8 and 9, the self-improving systems are those for which this limit equals one and the self-disimproving systems those for which it equals zero. An inspection of the table thus shows that the self-improving cases are precisely those with $d_{GG} = G$ and at least two of $\{d_{BB}, d_{BG}, d_{GB}\}$ equal to G , whereas the self-disimproving cases are exactly those with $d_{BB} = B$ and at least two of $\{d_{BG}, d_{GB}, d_{GG}\}$ equal to B . $\square$

Proposition 3 lets us distinguish three subsets of self-improving systems, which differ in their resident reputation equilibrium. The following lemma makes this partition explicit.

Lemma 4 (Partition of self-improving systems). *The set of self-improving systems $\mathcal{A}$ can be partitioned into the following three subsets, each characterized by a distinct resident reputation equilibrium in the absence of execution errors, $(g_{p,d}^*)_{\epsilon=0}$, as a function of the assignment error χ :*

1. The set $\mathcal{A}_{G^{**}G} = \mathcal{A}_{GBGG} \cup \mathcal{A}_{GGBG}$ of self-improving systems for which

$$d_{BB} = d_{GG} = G \text{ and } d_{BG} \neq d_{GB} \quad (70)$$

Table 2: Limiting resident reputation equilibrium $\lim_{\chi \rightarrow 0} \left(g_{p,d}^* \right)_{\epsilon=0}$ for each of the 16 possible assessment-bit combinations $(d_{BB}, d_{BG}, d_{GB}, d_{GG})$, in the absence of execution errors ($\epsilon = 0$) and in the limit of vanishing assignment error ($\chi \rightarrow 0$). Self-approving systems are those for which this limit equals one; self-disapproving systems those for which it equals zero.

d_{BB}	d_{BG}	d_{GB}	d_{GG}	$\lim_{\chi \rightarrow 0} \left(g_{p,d}^* \right)_{\epsilon=0}$
B	B	B	B	0
B	B	B	G	0
B	B	G	B	0
B	B	G	G	1/2
B	G	B	B	0
B	G	B	G	1/2
B	G	G	B	1/2
B	G	G	G	1
G	B	B	B	$\frac{3-\sqrt{5}}{2}$
G	B	B	G	1/2
G	B	G	B	1/2
G	B	G	G	1
G	G	B	B	1/2
G	G	B	G	1
G	G	G	B	$\frac{\sqrt{5}-1}{2}$
G	G	G	G	1

holds, characterized by

$$(g_{p,d}^*)_{\epsilon=0} = \frac{1 - \chi - \sqrt{\chi - \chi^2}}{1 - 2\chi}. \quad (71)$$

Here, $\mathcal{A}_{GBGG}$ denotes the subset with $(d_{BG}, d_{GB}) = (B, G)$ and $\mathcal{A}_{GGGG}$ the subset with $(d_{BG}, d_{GB}) = (G, B)$.

2. The set $\mathcal{A}_{BGGG}$ of self-approving systems for which

$$d_{BB} = B \text{ and } d_{BG} = d_{GB} = d_{GG} = G \quad (72)$$

holds, characterized by

$$(g_{p,d}^*)_{\epsilon=0} = \frac{1 - 4\chi + \sqrt{1 - 4\chi + 8\chi^2}}{2 - 4\chi}. \quad (73)$$

3. The set $\mathcal{A}_{GGGG}$ of self-approving systems for which

$$d_{BB} = d_{BG} = d_{GB} = d_{GG} = G \quad (74)$$

holds, characterized by

$$(g_{p,d}^*)_{\epsilon=0} = 1 - \chi. \quad (75)$$

Proof. The partition follows from Proposition 3: the systems satisfying its conditions are precisely those with $(d_{BB}, d_{BG}, d_{GB}, d_{GG})$ equal to (G, B, G, G) , (G, G, B, G) , (B, G, G, G) , or (G, G, G, G) , which group into the three subsets above. The explicit equilibria are obtained by substituting the corresponding assessment bits into Eqs. (21) and (22) at $\epsilon = 0$ (equivalently, into the $a(\chi)$, $b(\chi)$, and $c(\chi)$ of the proof of Proposition 3) and simplifying, without taking the $\chi \rightarrow 0$ limit. $\square$

3.2 Characterization of self-cooperative and righteous systems

The cooperation index in the absence of execution errors, $(\kappa_{p,d})_{\epsilon=0}$, and its limit when the assignment error tends to zero, $\lim_{\chi \rightarrow 0} (\kappa_{p,d})_{\epsilon=0}$, are important quantities in our analysis. The following lemma expresses these two quantities in terms of a simple quadratic function.

Lemma 5 (Cooperation index as a function of the reputation equilibrium). *Let $q_p : [0, 1] \rightarrow [0, 1]$ be defined by*

$$q_p(x) = p_{BB}(1 - x)^2 + (p_{BG} + p_{GB})x(1 - x) + p_{GG}x^2. \quad (76)$$

Then the cooperation index $(\kappa_{p,d})_{\epsilon=0}$ of a moral system (p, d) and its limit $\lim_{\chi \rightarrow 0} (\kappa_{p,d})_{\epsilon=0}$ can be written, respectively, as

$$(\kappa_{p,d})_{\epsilon=0} = q_p \left((g_{p,d}^*)_{\epsilon=0} \right), \quad (77)$$

and

$$\lim_{\chi \rightarrow 0} (\kappa_{p,d})_{\epsilon=0} = q_p \left(\lim_{\chi \rightarrow 0} (g_{p,d}^*)_{\epsilon=0} \right). \quad (78)$$

Proof. Evaluating the cooperation index (23) at $\epsilon = 0$ yields (77). Taking the limit of (77) as χ tends to zero and applying the algebraic limit theorem yields (78). $\square$

Table 3: The cooperation-index function $q_p(x)$ (76) for each of the 16 action rules p , where $x \in [0, 1]$ is the fraction of individuals with a good reputation. Actions and reputations follow the encoding C = G = 1, D = B = 0 (1).

p_{BB}	p_{BG}	p_{GB}	p_{GG}	$q_p(x)$
D	D	D	D	0
C	D	D	D	$1 - 2x + x^2$
D	C	D	D	$x - x^2$
C	C	D	D	$1 - x$
D	D	C	D	$x - x^2$
C	D	C	D	$1 - x$
D	C	C	D	$2x - 2x^2$
C	C	C	D	$1 - x^2$
D	D	D	C	x^2
C	D	D	C	$1 - 2x + 2x^2$
D	C	D	C	x
C	C	D	C	$1 - x + x^2$
D	D	C	C	x
C	D	C	C	$1 - x + x^2$
D	C	C	C	$2x - x^2$
C	C	C	C	1

Table 3 lists the cooperation-index function $q_p(x)$ (76) for each action rule p .

The following proposition provides simple necessary and sufficient conditions for a moral system to be self-cooperative. Its proof uses Lemma 5.

Proposition 4 (Necessary and sufficient conditions for a conditional system to be self-cooperative). *A conditional system is self-cooperative if and only if either*

- (i) *it is self-approving and $p_{GG} = C$ holds, or*
- (ii) *it is self-disapproving and $p_{BB} = C$ holds.*

Proof. By Lemma 5, we can write $\lim_{x \rightarrow 0} (\kappa_{p,d})_{\epsilon=0} = q_p(\gamma)$, where

$$\gamma = \lim_{x \rightarrow 0} (g_{p,d}^*)_{\epsilon=0} \quad (79)$$

and where q_p is given by (76). Substituting all possible choices of $(p_{BB}, p_{BG}, p_{GB}, p_{GG})$ into $q_p(\gamma)$ yields the 16 cases listed in Table 3.

Since p is conditional, there are three subcases to consider.

1. Suppose that $p_{BB} = p_{GG} = D$ holds. Then, $q_p(\gamma) = (p_{BG} + p_{GB}) \gamma(1 - \gamma)$ with $1 \leq p_{BG} + p_{GB} \leq 2$ (since p is conditional, p_{BG} and p_{GB} cannot both be D). Since $\gamma(1 - \gamma) \leq 1/4$ for $\gamma \in [0, 1]$, it follows that $\max_{0 \leq \gamma \leq 1} q_p(\gamma) \leq 1/2$. Hence, p can never be self-cooperative under d .
2. Suppose that $p_{GG} = C$ holds. Then, $q_p(1) = p_{GG} = 1$. Therefore, if (p, d) is self-approving (so that $\gamma = 1$ by Definition 8 and Eq. (79)), then it is self-cooperative. This proves the sufficient condition of part (i).

3. Suppose that $p_{BB} = C$ holds. Then, $q_p(0) = p_{BB} = 1$. Hence, if (p, d) is self-disapproving (so that $\gamma = 0$ by Definition 9 and Eq. (79)), then it is self-cooperative. This proves the sufficient condition of part (ii).

Finally, we prove the necessity of conditions (i)–(ii). Suppose (p, d) is self-cooperative, so that $q_p(\gamma) = 1$. If $0 < \gamma < 1$, all four weights $(1 - \gamma)^2$, $\gamma(1 - \gamma)$, $\gamma(1 - \gamma)$, and γ^2 in $q_p(\gamma)$ are strictly positive, so $q_p(\gamma) = 1$ would require $p_{BB} = p_{GG} = C$ and $p_{BG} = p_{GB} = C$, that is, $p = ALLC$; since p is conditional, this is impossible, and hence $q_p(\gamma) < 1$, a contradiction. Therefore $\gamma \in \{0, 1\}$, so (p, d) is self-approving or self-disapproving. If $\gamma = 1$, then $p_{GG} = q_p(1) = 1$, so $p_{GG} = C$ and (i) holds; if $\gamma = 0$, then $p_{BB} = q_p(0) = 1$, so $p_{BB} = C$ and (ii) holds. This concludes the proof. $\square$

Corollary 2 (Characterization of righteous systems). *Let (p, d) be self-approving. Then (p, d) is self-cooperative (and hence also righteous) if and only if $p_{GG} = C$.*

Proof. Being self-approving, (p, d) is not self-disapproving, so condition (ii) of Proposition 4 cannot hold. Hence (p, d) is self-cooperative if and only if condition (i) holds, that is, if and only if $p_{GG} = C$. Righteousness then follows from Definition 11, since (p, d) is self-approving by assumption. $\square$

Proposition 4 shows that the self-cooperative systems form two disjoint classes. The first is the set of self-approving systems satisfying $p_{GG} = C$; by Corollary 2, these coincide with the righteous systems of Definition 11. The second is the set of self-disapproving systems satisfying $p_{BB} = C$.

Together with Proposition 3, Corollary 2 lets us simplify the characterization of the set $\mathcal{R}$ of righteous systems given in Eq. (42) to

$$\mathcal{R} = \{(p, d) \mid (p, d) \in C, p_{GG} = C, d_{GG} = G, \text{ and } d_{BB} + d_{BG} + d_{GB} \geq 2\}. \quad (80)$$

This characterization also lets us count the righteous systems, a figure reported in Fig. 4.

Proposition 5 (Cardinality of $\mathcal{R}$). *The set of righteous systems contains $|\mathcal{R}| = 114,688$ moral systems.*

Proof. An assessment rule d is a function of $2^4 = 16$ argument combinations, hence a string of 16 bits; a condition that leaves k of these bits unconstrained is satisfied by 2^k assessment rules. We count the admissible systems by multiplying, over action rules and constrained bits, the number of choices each leaves open. By Eq. (80) a righteous system is conditional and has $p_{GG} = C$. Fixing $p_{GG} = C$ leaves $2^3 = 8$ strategies; one of them, the unconditional cooperator ALLC ($p = CCCC$), is excluded because a righteous system is conditional (Definition 11), leaving 7 admissible action rules. Each moreover satisfies $d_{GG} = G$ together with $d_{BB} + d_{BG} + d_{GB} \geq 2$, which constrain four distinct bits of d : d_{GG} is fixed to G, while of the eight assignments of (d_{BB}, d_{BG}, d_{GB}) only the four with at least two G's are admissible. The remaining 12 bits are free, so

$$|\mathcal{R}| = 7 \cdot 4 \cdot 2^{12} = 114,688.$$

$\square$

3.3 Righteousness and reputational dominance of consistent systems

Every consistent system is self-approving, as the following proposition records.

Proposition 6 (Consistent systems are self-approving). *If (p, d) is consistent, then it lies in $\mathcal{A}_{GGGG}$ and is, in particular, self-approving. That is, $Q \subset \mathcal{A}_{GGGG} \subset \mathcal{A}$.*

Proof. By condition (43a) of Definition 13, a consistent system satisfies $d_{GG} = G$ and $d_{BB} = d_{BG} = d_{GB} = G$, so that $d_{BB} + d_{BG} + d_{GB} = 3 \geq 2$. By Proposition 3, it is therefore self-approving. Since, in addition, all four assessment bits equal G, it lies in the subset $\mathcal{A}_{GGGG}$ (item 3 of Lemma 4); hence $Q \subset \mathcal{A}_{GGGG} \subset \mathcal{A}$. $\square$

By Definition 14, a consistent discriminating system (p, d) is one whose action rule p is discriminating and consistent with d .

Proposition 7 (Consistent discriminating systems are righteous). *If (p, d) is consistent discriminating, then it is righteous. That is, $Q_{\mathcal{D}} \subset \mathcal{R}$.*

Proof. Since p is discriminating, it is conditional and satisfies $p_{GG} = C$ (Definitions 2 and 3). Being consistent, (p, d) is self-approving (Proposition 6); together with $p_{GG} = C$, Proposition 4(i) then makes it self-cooperative. A system that is both self-approving and self-cooperative is righteous (Definition 11), so $Q_{\mathcal{D}} \subset \mathcal{R}$. $\square$

The remainder of this subsection characterizes consistency in reputational terms. The intuition is that a consistent assessment rule labels as good exactly those who follow the resident prescription, so no alternative action rule should be able to match the resident's reputation under d . The next two propositions make this claim precise at $\epsilon = 0$: together, they show that a system is consistent if and only if its resident action rule attains a strictly higher reputation than every mutant action rule.

Proposition 8 (Reputational dominance of consistent systems). *Let $\epsilon = 0$ and $\chi \in (0, 1/2)$. If $(p, d) \in Q$, then for every mutant action rule $\tilde{p} \neq p$,*

$$h_{\tilde{p}, p, d}^* < g_{p, d}^*.$$

Proof. By Definition 13, $d_{BB} = d_{BG} = d_{GB} = d_{GG} = G$. For $\epsilon = 0$, Eq. (19) reduces to $\delta_{r_1 r_2} = \chi + (1 - 2\chi) d_{r_1 r_2} = 1 - \chi$ for all $(r_1, r_2) \in \{B, G\}^2$. Substituting into Eq. (21) yields $g_{p, d}^* = 1 - \chi$, in agreement with Lemma 4, item 3.

Fix a mutant $\tilde{p} \neq p$ and, for each $(r_1, r_2) \in \{B, G\}^2$, let $b_{r_1 r_2} = 1$ if $\tilde{p}(r_1, r_2) = p(r_1, r_2)$ and $b_{r_1 r_2} = 0$ otherwise. Because (p, d) is consistent, whenever $\tilde{p}(r_1, r_2) \neq p(r_1, r_2)$ we have $\tilde{d}_{r_1 r_2} = \bar{d}_{r_1 r_2} = B$ (by condition (43b)); and whenever $\tilde{p}(r_1, r_2) = p(r_1, r_2)$ we have $\tilde{d}_{r_1 r_2} = d_{r_1 r_2} = G$ (by condition (43a)). Hence $\tilde{d}_{r_1 r_2} = G$ if and only if $b_{r_1 r_2} = 1$, and, for $\epsilon = 0$, Eq. (24) reduces to

$$\tilde{\delta}_{r_1 r_2} = \chi + (1 - 2\chi) b_{r_1 r_2}. \quad (81)$$

The mutant reputation equilibrium $h_{\tilde{p}, p, d}^*$, obtained by setting $\dot{h}_{\tilde{p}, p, d} = 0$ in Eq. (25), satisfies the fixed-point equation

$$h_{\tilde{p}, p, d}^* = \alpha + h_{\tilde{p}, p, d}^* (\beta - \alpha), \quad (82)$$

with

$$\alpha := g_{p, d}^* \tilde{\delta}_{BG} + (1 - g_{p, d}^*) \tilde{\delta}_{BB}, \quad (83a)$$

$$\beta := g_{p, d}^* \tilde{\delta}_{GG} + (1 - g_{p, d}^*) \tilde{\delta}_{GB}. \quad (83b)$$

Since $\tilde{\delta}_{r_1 r_2} \in \{\chi, 1 - \chi\}$ for every $(r_1, r_2) \in \{B, G\}^2$, and since α and β are convex combinations of these values (the weights $g_{p, d}^*$ and $1 - g_{p, d}^*$ being nonnegative and summing to one), both α

and $1 - \beta$ lie in $[\chi, 1 - \chi]$, so $1 - \beta + \alpha \geq 2\chi > 0$. Solving Eq. (82) for $h_{\tilde{p},p,d}^*$ (legitimate since $1 - \beta + \alpha > 0$) gives

$$h_{\tilde{p},p,d}^* = \frac{\alpha}{1 - \beta + \alpha},$$

in agreement with Eq. (26), and a direct computation gives

$$g_{p,d}^* - h_{\tilde{p},p,d}^* = \frac{(1 - \chi)(1 - \beta) - \chi \alpha}{1 - \beta + \alpha}.$$

To expand the numerator, let $\rho_{r_1 r_2}$ denote the equilibrium probability that a focal player has reputation r_1 and the co-player has reputation r_2 ; these are the reputation-pair weights $(g_{p,d}^*)^2$, $g_{p,d}^*(1 - g_{p,d}^*)$, and $(1 - g_{p,d}^*)^2$ that already appear in Eqs. (20) and (28). With $g_{p,d}^* = 1 - \chi$,

$$\rho_{GG} = (1 - \chi)^2, \quad \rho_{GB} = \rho_{BG} = \chi(1 - \chi), \quad \rho_{BB} = \chi^2,$$

which satisfy

$$\sum_{r_1, r_2 \in \{B, G\}} \rho_{r_1 r_2} = [(1 - \chi) + \chi]^2 = 1. \quad (84)$$

Within α and β the $\tilde{\delta}_{r_1 r_2}$ are weighted by $g_{p,d}^*$ and $1 - g_{p,d}^*$, and the outer coefficients $1 - \chi = g_{p,d}^*$ and $\chi = 1 - g_{p,d}^*$ match these; hence in $(1 - \chi)\beta + \chi \alpha$ the coefficient multiplying each $\tilde{\delta}_{r_1 r_2}$ is exactly the product $\rho_{r_1 r_2}$. Collecting the four terms by reputation pair gives

$$(1 - \chi)\beta + \chi \alpha = \sum_{r_1, r_2 \in \{B, G\}} \rho_{r_1 r_2} \tilde{\delta}_{r_1 r_2}. \quad (85)$$

Substituting $\tilde{\delta}_{r_1 r_2} = \chi + (1 - 2\chi)b_{r_1 r_2}$ from Eq. (81), the numerator becomes

$$\begin{aligned} (1 - \chi)(1 - \beta) - \chi \alpha &= (1 - \chi) - [(1 - \chi)\beta + \chi \alpha] \\ &= (1 - \chi) - \sum_{r_1, r_2 \in \{B, G\}} \rho_{r_1 r_2} [\chi + (1 - 2\chi)b_{r_1 r_2}] \\ &= (1 - \chi) - \chi \sum_{r_1, r_2 \in \{B, G\}} \rho_{r_1 r_2} - (1 - 2\chi) \sum_{r_1, r_2 \in \{B, G\}} \rho_{r_1 r_2} b_{r_1 r_2} \\ &= (1 - \chi) - \chi - (1 - 2\chi) \sum_{r_1, r_2 \in \{B, G\}} \rho_{r_1 r_2} b_{r_1 r_2} \\ &= (1 - 2\chi) \left(\sum_{r_1, r_2 \in \{B, G\}} \rho_{r_1 r_2} - \sum_{r_1, r_2 \in \{B, G\}} \rho_{r_1 r_2} b_{r_1 r_2} \right) \\ &= (1 - 2\chi) \sum_{r_1, r_2 \in \{B, G\}} \rho_{r_1 r_2} (1 - b_{r_1 r_2}) \\ &= (1 - 2\chi) \left\{ (1 - \chi)^2 (1 - b_{GG}) + \chi(1 - \chi)(1 - b_{GB}) \right. \\ &\quad \left. + \chi(1 - \chi)(1 - b_{BG}) + \chi^2(1 - b_{BB}) \right\}, \end{aligned} \quad (86)$$

where the second equality uses Eq. (85); the fourth and fifth equalities both invoke the identity (84) — the fourth to replace $\chi \sum_{r_1, r_2} \rho_{r_1 r_2}$ by χ , the fifth to reinstate $1 = \sum_{r_1, r_2} \rho_{r_1 r_2}$ (after combining $1 - \chi - \chi = 1 - 2\chi$). Each $(1 - b_{r_1 r_2}) \in \{0, 1\}$, the weights $\rho_{r_1 r_2}$ are strictly positive for $\chi \in (0, 1/2)$, and at least one $(1 - b_{r_1 r_2}) = 1$ since $\tilde{p} \neq p$. Hence the right-hand side of Eq. (86) is strictly positive, and $g_{p,d}^* - h_{\tilde{p},p,d}^* > 0$. $\square$

Proposition 9 (Reputational non-dominance of inconsistent systems). *Let $\epsilon = 0$ and $\chi \in (0, 1/2)$. If $(p, d) \notin Q$, then there exists $\tilde{p} \neq p$ such that*

$$h_{\tilde{p},p,d}^* \geq g_{p,d}^*.$$

Proof. Because $(p, d) \notin Q$, Definition 13 guarantees that at least one pair $(\mathbf{d}_{r_1 r_2}, \bar{\mathbf{d}}_{r_1 r_2})$ differs from (G, B) . Throughout, (r_1, r_2) ranges over the four reputation pairs in $\{B, G\}^2$. We distinguish two mutually exclusive and exhaustive cases.

1. Suppose some pair (r_1^*, r_2^*) has $\mathbf{d}_{r_1^* r_2^*} = \bar{\mathbf{d}}_{r_1^* r_2^*}$. Let $\tilde{p}$ be the single-bit flip of p at (r_1^*, r_2^*) : $\tilde{p}(r_1^*, r_2^*) = \neg p(r_1^*, r_2^*)$ and $\tilde{p} = p$ elsewhere, so that $\tilde{p} \neq p$. We claim that $\tilde{\mathbf{d}}_{r_1 r_2} = \mathbf{d}_{r_1 r_2}$ at every pair. At (r_1^*, r_2^*) , the flip gives $\tilde{\mathbf{d}}_{r_1^* r_2^*} = \bar{\mathbf{d}}_{r_1^* r_2^*}$ (by the shorthands (16e) and (16f)), which equals $\mathbf{d}_{r_1^* r_2^*}$ by the case hypothesis; at every other pair, $\tilde{p} = p$, so $\tilde{\mathbf{d}}_{r_1 r_2} = \mathbf{d}_{r_1 r_2}$ trivially. At $\epsilon = 0$, Eqs. (19) and (24) reduce to $\delta_{r_1 r_2} = \chi + (1 - 2\chi) \mathbf{d}_{r_1 r_2}$ and $\tilde{\delta}_{r_1 r_2} = \chi + (1 - 2\chi) \tilde{\mathbf{d}}_{r_1 r_2}$, so the equalities $\tilde{\mathbf{d}}_{r_1 r_2} = \mathbf{d}_{r_1 r_2}$ yield $\tilde{\delta}_{r_1 r_2} = \delta_{r_1 r_2}$ at every pair. Substituting these equalities into the mutant reputation dynamics (25) and evaluating its right-hand side at $h_{\tilde{p},p,d}^* = g_{p,d}^*$ gives

$$\delta_{GG} (g_{p,d}^*)^2 + (\delta_{BG} + \delta_{GB}) g_{p,d}^* (1 - g_{p,d}^*) + \delta_{BB} (1 - g_{p,d}^*)^2 - g_{p,d}^*.$$

This is exactly the right-hand side of the resident dynamics (20) at $g_{p,d} = g_{p,d}^*$, which vanishes by the definition of $g_{p,d}^*$. Thus $h_{\tilde{p},p,d}^* = g_{p,d}^*$ is an equilibrium of (25), and by the uniqueness of the reputation equilibria (Eqs. (21) and (26)), $h_{\tilde{p},p,d}^* = g_{p,d}^*$.

2. Suppose instead that $\mathbf{d}_{r_1 r_2} \neq \bar{\mathbf{d}}_{r_1 r_2}$ at every pair, so that each $(\mathbf{d}_{r_1 r_2}, \bar{\mathbf{d}}_{r_1 r_2})$ is either (G, B) or (B, G) . Were all four pairs (G, B) , Definition 13 would make (p, d) consistent, contradicting the hypothesis; hence the set

$$S := \{(r_1, r_2) : (\mathbf{d}_{r_1 r_2}, \bar{\mathbf{d}}_{r_1 r_2}) = (B, G)\}$$

is non-empty. Let $\tilde{p}$ flip p on S and agree with p elsewhere, so that $\tilde{p} \neq p$. Then $\tilde{\mathbf{d}}_{r_1 r_2} = G$ at every pair: for $(r_1, r_2) \in S$, $\tilde{\mathbf{d}}_{r_1 r_2} = \bar{\mathbf{d}}_{r_1 r_2} = G$; for $(r_1, r_2) \notin S$, $\tilde{\mathbf{d}}_{r_1 r_2} = \mathbf{d}_{r_1 r_2} = G$, since $(\mathbf{d}_{r_1 r_2}, \bar{\mathbf{d}}_{r_1 r_2}) = (G, B)$. At $\epsilon = 0$, Eq. (24) therefore reduces to $\tilde{\delta}_{r_1 r_2} = \chi + (1 - 2\chi) \tilde{\mathbf{d}}_{r_1 r_2} = 1 - \chi$ at every pair. Substituting into Eq. (26), we find that the coefficients of $g_{p,d}^*$ in the numerator, $\tilde{\delta}_{BG} - \tilde{\delta}_{BB}$, and in the denominator, $\tilde{\delta}_{BG} + \tilde{\delta}_{GB} - \tilde{\delta}_{BB} - \tilde{\delta}_{GG}$, both vanish; the dependence on $g_{p,d}^*$ drops out, leaving $h_{\tilde{p},p,d}^* = 1 - \chi$. It remains to compare this with $g_{p,d}^*$. Setting $\dot{g}_{p,d} = 0$ in (20) expresses $g_{p,d}^*$ as a convex combination of the $\delta_{r_1 r_2}$. At $\epsilon = 0$, each $\delta_{r_1 r_2} = \chi + (1 - 2\chi) \mathbf{d}_{r_1 r_2}$ equals χ (if $\mathbf{d}_{r_1 r_2} = B$) or $1 - \chi$ (if $\mathbf{d}_{r_1 r_2} = G$), and $\chi < 1 - \chi$ for $\chi \in (0, 1/2)$, so every $\delta_{r_1 r_2} \leq 1 - \chi$; hence $g_{p,d}^* \leq 1 - \chi$. Therefore $h_{\tilde{p},p,d}^* = 1 - \chi \geq g_{p,d}^*$. $\square$

Combining Propositions 8 and 9 yields the reputational characterization of consistency anticipated at the start of this discussion.

Corollary 3 (Reputational characterization of consistency). *Let $\epsilon = 0$ and $\chi \in (0, 1/2)$. A system (p, d) is consistent if and only if $h_{\tilde{p},p,d}^* < g_{p,d}^*$ for every mutant action rule $\tilde{p} \neq p$.*

Proof. Proposition 8 gives the forward direction. The contrapositive of the reverse direction is Proposition 9: if $(p, d) \notin Q$, then some $\tilde{p} \neq p$ satisfies $h_{\tilde{p},p,d}^* \geq g_{p,d}^*$. $\square$

3.4 Righteousness of the leading eight

The leading eight (60) have action rule $p = *CDC$. Together with the shorthand (16d) and the leading-eight assessment rule (48), this implies

$$d_{BG} = d_{GB} = d_{GG} = G. \quad (87)$$

These values already place the leading eight among the righteous systems.

Proposition 10 (Leading-eight systems are righteous). *Every leading-eight moral system is righteous; that is, $\mathcal{L} \subset \mathcal{R}$. Moreover, those in groups I and II belong to $\mathcal{A}_{GGGG}$, while those in group III belong to $\mathcal{A}_{BGGG}$.*

Proof. By Eq. (87), $d_{GG} = G$ and at least two of $\{d_{BB}, d_{BG}, d_{GB}\}$ equal G , so the leading eight are self-approving (Proposition 3). Since $p_{GG} = C$, they are also self-cooperative (Corollary 2), and hence righteous (Definition 11); thus $\mathcal{L} \subset \mathcal{R}$. For the finer placement, Eqs. (49)–(57) give $d_{BB} = G$ for groups I and II and $d_{BB} = B$ for group III. Combined with Eq. (87), the former therefore have all four assessment bits equal to G and lie in $\mathcal{A}_{GGGG}$, while the latter lie in $\mathcal{A}_{BGGG}$. $\square$

3.5 Universal stability within the set of righteous systems

Many moral systems of interest are righteous, notably the consistent discriminating systems (Proposition 7) and the leading-eight systems (Proposition 10). It is therefore natural to ask which righteous systems are universally stable (Definition 12). We show that a righteous system is universally stable if and only if it is a contrite consistent discriminating system; that is, the universally stable systems in the righteous set $\mathcal{R}$ (Eq. (80)) are exactly $\mathcal{Q}_{CDISC}$ (Eq. (45)).

We establish necessity by eliminating systems one class of invader at a time, producing a strictly decreasing chain $\mathcal{R} \supset \mathcal{R}' \supset \mathcal{R}'' \supset \mathcal{R}''' = \mathcal{Q}_{CDISC}$: we discard in turn the systems invaded by ALLD (leaving $\mathcal{R}'$; Section 3.5.2), by ALLC (leaving $\mathcal{R}''$; Section 3.5.3), and by some discriminating strategy (leaving $\mathcal{R}''' = \mathcal{Q}_{CDISC}$; Section 3.5.4). Sufficiency then follows by verifying that every system in $\mathcal{Q}_{CDISC}$ resists invasion by all alternative strategies (Section 3.5.5). Fig. 4 maps this chain and locates the consistent discriminating systems and the leading eight within it.

3.5.1 Preliminaries

Reputation equilibria. Given a resident system $(p, d) \in \mathcal{R}$ and a mutant strategy $\tilde{p} \neq p$, we write

$$g(\chi) := (g_{p,d}^*)_{\epsilon=0} \quad (88)$$

and

$$h(\chi) := \left(h_{\tilde{p},p,d}^* \right)_{\epsilon=0} \quad (89)$$

for the resident and mutant reputation equilibria at $\epsilon = 0$, as functions of χ .

Since a righteous resident $(p, d) \in \mathcal{R}$ is self-approving (Definition 11), it belongs to exactly one of the three classes $\mathcal{A}_{G^{**}G}$, $\mathcal{A}_{BGGG}$, and $\mathcal{A}_{GGGG}$ that partition the self-approving systems (Lemma 4); accordingly, $g(\chi)$ is given (Eqs. (71)–(75)) by

$$g(\chi) = \begin{cases} \frac{1-\chi-\sqrt{\chi-\chi^2}}{1-2\chi} & \text{if } (p, d) \in \mathcal{A}_{G^{**}G} \\ \frac{1-4\chi+\sqrt{1-4\chi+8\chi^2}}{2-4\chi} & \text{if } (p, d) \in \mathcal{A}_{BGGG} \\ 1-\chi & \text{if } (p, d) \in \mathcal{A}_{GGGG} \end{cases} \quad (90)$$

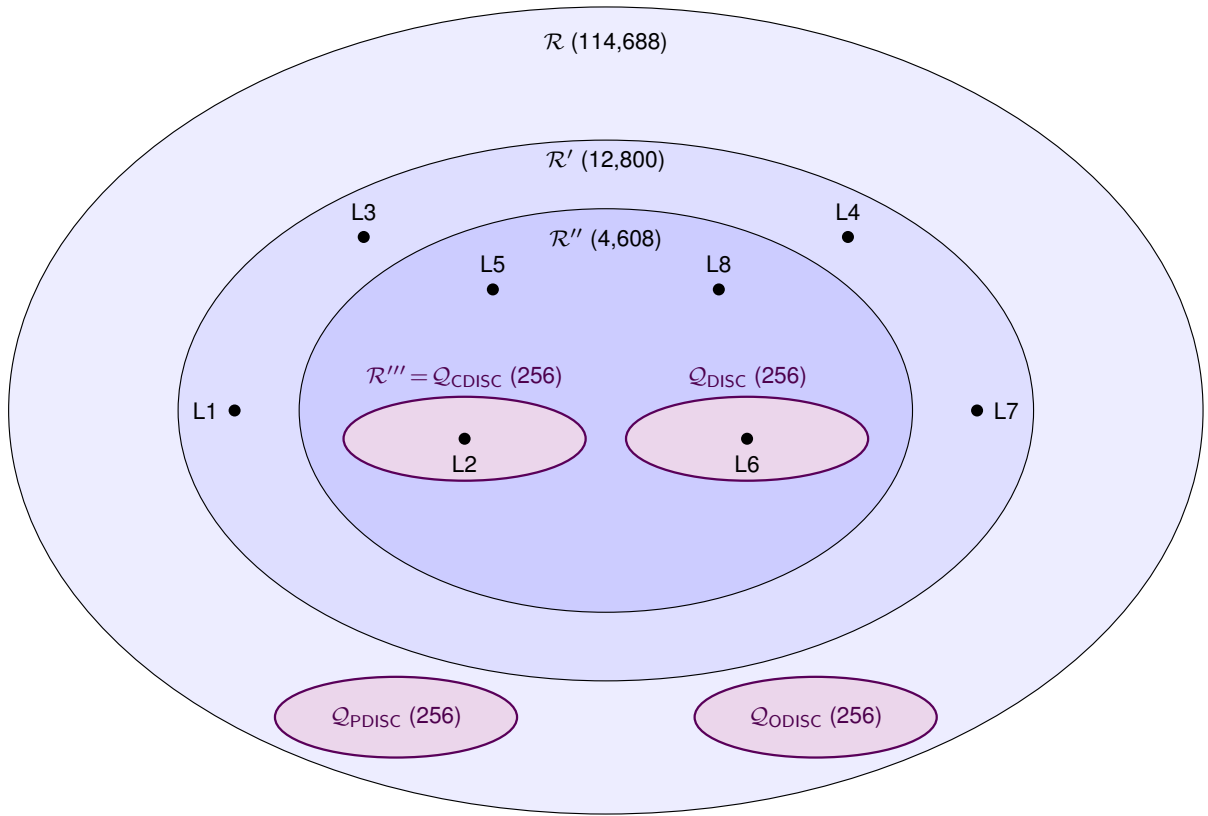


Figure 4: The nested righteous sets produced by successively removing the systems invaded by ALLD, by ALLC, and by a discriminating strategy: $\mathcal{R} \supset \mathcal{R}' \supset \mathcal{R}'' \supset \mathcal{R}''' = \mathcal{Q}_{\text{CDISC}}$. This leaves $\mathcal{Q}_{\text{CDISC}}$ as the only candidate for universal stability; a final check that it resists every remaining invader shows that the contrite consistent discriminating systems $\mathcal{Q}_{\text{CDISC}}$ are the only universally stable righteous systems (Section 3.5). The diagram is schematic (not to scale), but sets of equal cardinality are drawn at equal size, and the parenthetical figures give the number of moral systems in each set. The four consistent discriminating families appear as equal purple ellipses; the leading eight, whose assessment rules are third-order, are single systems shown as dots, with L2 (CDISC, Consistent Standing) in $\mathcal{Q}_{\text{CDISC}}$, and L6 (DISC, Stern Judging) in $\mathcal{Q}_{\text{DISC}}$.

Substituting $\epsilon = 0$ into the mutant equilibrium (26) — and thus into the assessment probabilities $\tilde{\delta}_{r_1 r_2}$ (24) on which it depends — and simplifying gives

$$h(\chi) = \frac{\chi + (1 - 2\chi)\tilde{d}_{BB} + (1 - 2\chi)(\tilde{d}_{BG} - \tilde{d}_{BB})g(\chi)}{1 + (1 - 2\chi)(\tilde{d}_{BB} - \tilde{d}_{GB}) + (1 - 2\chi)(\tilde{d}_{BG} + \tilde{d}_{GB} - \tilde{d}_{BB} - \tilde{d}_{GG})g(\chi)},$$

which depends on the triple $(p, d, \tilde{p})$ only through $g(\chi)$ (90) and the four mutant assessment bits $\tilde{d}_{BB}$, $\tilde{d}_{BG}$, $\tilde{d}_{GB}$, and $\tilde{d}_{GG}$ (16e).

Expected payoffs. We similarly write

$$w(\chi) := (w_{p,d})_{\epsilon=0} \quad (91)$$

and

$$\tilde{w}(\chi) := (\tilde{w}_{\tilde{p},p,d})_{\epsilon=0} \quad (92)$$

for the expected payoffs to a resident and a mutant, evaluated at the reputation equilibrium and at $\epsilon = 0$, as functions of χ . Substituting $\epsilon = 0$ into Eq. (28), we obtain

$$\begin{aligned} w(\chi) &= (1 - g(\chi))^2 (\omega_{BB})_{\epsilon=0} \\ &\quad + g(\chi)(1 - g(\chi)) ((\omega_{BG})_{\epsilon=0} + (\omega_{GB})_{\epsilon=0}) \\ &\quad + g(\chi)^2 (\omega_{GG})_{\epsilon=0}, \end{aligned} \quad (93)$$

where, by Eq. (27) at $\epsilon = 0$,

$$(\omega_{r_1 r_2})_{\epsilon=0} = v(\mathbf{p}_{r_1 r_2}, \mathbf{p}_{r_2 r_1}).$$

Likewise, substituting $\epsilon = 0$ into Eq. (30), we obtain

$$\begin{aligned} \tilde{w}(\chi) &= (1 - h(\chi))(1 - g(\chi)) (\tilde{\omega}_{BB})_{\epsilon=0} + (1 - h(\chi))g(\chi) (\tilde{\omega}_{BG})_{\epsilon=0} \\ &\quad + h(\chi)(1 - g(\chi)) (\tilde{\omega}_{GB})_{\epsilon=0} + h(\chi)g(\chi) (\tilde{\omega}_{GG})_{\epsilon=0}, \end{aligned} \quad (94)$$

where, by Eq. (29) at $\epsilon = 0$,

$$(\tilde{\omega}_{r_1 r_2})_{\epsilon=0} = v(\tilde{\mathbf{p}}_{r_1 r_2}, \mathbf{p}_{r_2 r_1}).$$

Thus, $w(\chi)$ is determined by $g(\chi)$ and the resident payoffs $(\omega_{BB})_{\epsilon=0}$, $(\omega_{BG})_{\epsilon=0}$, $(\omega_{GB})_{\epsilon=0}$, $(\omega_{GG})_{\epsilon=0}$, and $\tilde{w}(\chi)$ additionally by $h(\chi)$ and the mutant payoffs $(\tilde{\omega}_{BB})_{\epsilon=0}$, $(\tilde{\omega}_{BG})_{\epsilon=0}$, $(\tilde{\omega}_{GB})_{\epsilon=0}$, $(\tilde{\omega}_{GG})_{\epsilon=0}$; each of these payoffs equals one of R, S, T, P , according to the actions the focal player and co-player take.

Stability criterion. We assess whether p resists invasion by $\tilde{p}$ under d by expanding $w(\chi)$ and $\tilde{w}(\chi)$ as $\chi \rightarrow 0$ and checking whether the leading term of

$$\Delta w(\chi) := w(\chi) - \tilde{w}(\chi) \quad (95)$$

is positive, so that condition (31) holds at $\epsilon = 0$ for sufficiently small χ . Universal stability of p under d (Definition 12) then requires this to hold for all $\tilde{p} \neq p$ across all cooperative dilemmas we study. Since $g(\chi)$ takes a different form on each of $\mathcal{A}_{G^{**}G}$, $\mathcal{A}_{BG^{**}G}$, and $\mathcal{A}_{GG^{**}G}$ (Eq. (90)), we treat these classes separately.

Resident system. By Eq. (80), a resident system $(p, d) \in \mathcal{R}$ satisfies

$$p_{GG} = C \quad (96)$$

and

$$d_{GG} = G, \quad (97a)$$

$$d_{BB} + d_{BG} + d_{GB} \geq 2. \quad (97b)$$

Recalling the shorthand (16d), $d_{GG} = d(G, G, p_{GG}, p_{GG})$, so Eqs. (96) and (97a) together imply $d_{GGCC} = G$.

3.5.2 Invasion attempts by ALLD

For the mutant $\tilde{p} = \text{ALLD}$,

$$\tilde{p}_{BB} = \tilde{p}_{BG} = \tilde{p}_{GB} = \tilde{p}_{GG} = D. \quad (98)$$

We seek the subset $\mathcal{R}' \subseteq \mathcal{R}$ of righteous systems that resist such invasion across cooperative dilemmas. This subset is characterized by the following two propositions, according to whether $d_{GB} = G$ or $d_{GB} = B$; both require $p_{GB} = D$ and differ only in condition (ii). We prove them below by a single argument.

Proposition 11 (Stability against ALLD for righteous systems with $d_{GB} = G$). *Let $\epsilon = 0$ and $\chi \rightarrow 0$. A righteous system (p, d) with $d_{GB} = G$ resists invasion attempts by ALLD across cooperative dilemmas if and only if*

(i) $p_{GB} = D$, and

(ii) $d_{BGDD} = d_{GGDC} = B$.

Proposition 12 (Stability against ALLD for righteous systems with $d_{GB} = B$). *Let $\epsilon = 0$ and $\chi \rightarrow 0$. A righteous system (p, d) with $d_{GB} = B$ resists invasion attempts by ALLD across cooperative dilemmas if and only if*

(i) $p_{GB} = D$, and

(ii) $d_{BGDD} = B$ and either $d_{GGDC} = B$ or $d(B, B, D, p_{BB}) = B$.

Proof of Propositions 11 and 12. Since any righteous system is self-improving (Definitions 8 and 11), $\lim_{\chi \rightarrow 0} g(\chi) = 1$, and hence

$$\lim_{\chi \rightarrow 0} w(\chi) = (\omega_{GG})_{\epsilon=0} = v(p_{GG}, p_{GG}) = v_{CC} = R. \quad (99)$$

Since $w(\chi) \rightarrow R$ as $\chi \rightarrow 0$, the leading term of $\Delta w(\chi)$ has the sign of $R - \lim_{\chi \rightarrow 0} \tilde{w}(\chi)$: the resident resists ALLD when $\lim_{\chi \rightarrow 0} \tilde{w}(\chi) < R$ and is invaded when $\lim_{\chi \rightarrow 0} \tilde{w}(\chi) > R$.

As $\chi \rightarrow 0$, $g(\chi) \rightarrow 1$, so the terms of $\tilde{w}(\chi)$ (94) carrying a factor $1 - g(\chi)$ vanish, leaving

$$\lim_{\chi \rightarrow 0} \tilde{w}(\chi) = (1 - h_0) (\tilde{\omega}_{BG})_{\epsilon=0} + h_0 (\tilde{\omega}_{GG})_{\epsilon=0},$$

where

$$h_0 := \lim_{\chi \rightarrow 0} h(\chi).$$

Table 4: $\lim_{\chi \rightarrow 0} \tilde{w}(\chi)$ for invasion attempts by ALLD against righteous resident systems $(p, d) \in \mathcal{R}$. Rows enumerate all combinations of the mutant assessment bits ($\tilde{d}_{BB}, \tilde{d}_{BG}, \tilde{d}_{GB}, \tilde{d}_{GG}$). The last two columns give $\lim_{\chi \rightarrow 0} \tilde{w}(\chi)$ separately for the two mutually exclusive resident classes $(p, d) \in \mathcal{A}_{G^{**}G}$ and $(p, d) \in \mathcal{A}_{BGGG} \cup \mathcal{A}_{GGGG}$, in terms of $(\tilde{w}_{BG})_{\epsilon=0}$ and $(\tilde{w}_{GG})_{\epsilon=0}$ defined in Eqs. (100).

$\tilde{d}_{BB}$	$\tilde{d}_{BG}$	$\tilde{d}_{GB}$	$\tilde{d}_{GG}$	$(p, d) \in \mathcal{A}_{G^{**}G}$	$(p, d) \in \mathcal{A}_{BGGG} \cup \mathcal{A}_{GGGG}$
B	B	B	B	$(\tilde{w}_{BG})_{\epsilon=0}$	$(\tilde{w}_{BG})_{\epsilon=0}$
B	B	B	G	$(\tilde{w}_{BG})_{\epsilon=0}$	$\frac{2}{3}(\tilde{w}_{BG})_{\epsilon=0} + \frac{1}{3}(\tilde{w}_{GG})_{\epsilon=0}$
B	B	G	B	$(\tilde{w}_{BG})_{\epsilon=0}$	$(\tilde{w}_{BG})_{\epsilon=0}$
B	B	G	G	$\frac{1}{2}(\tilde{w}_{BG})_{\epsilon=0} + \frac{1}{2}(\tilde{w}_{GG})_{\epsilon=0}$	$\frac{1}{2}(\tilde{w}_{BG})_{\epsilon=0} + \frac{1}{2}(\tilde{w}_{GG})_{\epsilon=0}$
B	G	B	B	$\frac{1}{2}(\tilde{w}_{BG})_{\epsilon=0} + \frac{1}{2}(\tilde{w}_{GG})_{\epsilon=0}$	$\frac{1}{2}(\tilde{w}_{BG})_{\epsilon=0} + \frac{1}{2}(\tilde{w}_{GG})_{\epsilon=0}$
B	G	B	G	$(\tilde{w}_{GG})_{\epsilon=0}$	$(\tilde{w}_{GG})_{\epsilon=0}$
B	G	G	B	$\frac{1}{2}(\tilde{w}_{BG})_{\epsilon=0} + \frac{1}{2}(\tilde{w}_{GG})_{\epsilon=0}$	$\frac{1}{2}(\tilde{w}_{BG})_{\epsilon=0} + \frac{1}{2}(\tilde{w}_{GG})_{\epsilon=0}$
B	G	G	G	$(\tilde{w}_{GG})_{\epsilon=0}$	$(\tilde{w}_{GG})_{\epsilon=0}$
G	B	B	B	$(\tilde{w}_{BG})_{\epsilon=0}$	$(\tilde{w}_{BG})_{\epsilon=0}$
G	B	B	G	$\frac{1}{2}(\tilde{w}_{BG})_{\epsilon=0} + \frac{1}{2}(\tilde{w}_{GG})_{\epsilon=0}$	$\frac{1}{2}(\tilde{w}_{BG})_{\epsilon=0} + \frac{1}{2}(\tilde{w}_{GG})_{\epsilon=0}$
G	B	G	B	$(\tilde{w}_{BG})_{\epsilon=0}$	$(\tilde{w}_{BG})_{\epsilon=0}$
G	B	G	G	$(\tilde{w}_{GG})_{\epsilon=0}$	$\frac{1}{3}(\tilde{w}_{BG})_{\epsilon=0} + \frac{2}{3}(\tilde{w}_{GG})_{\epsilon=0}$
G	G	B	B	$\frac{1}{2}(\tilde{w}_{BG})_{\epsilon=0} + \frac{1}{2}(\tilde{w}_{GG})_{\epsilon=0}$	$\frac{1}{2}(\tilde{w}_{BG})_{\epsilon=0} + \frac{1}{2}(\tilde{w}_{GG})_{\epsilon=0}$
G	G	B	G	$(\tilde{w}_{GG})_{\epsilon=0}$	$(\tilde{w}_{GG})_{\epsilon=0}$
G	G	G	B	$\frac{1}{2}(\tilde{w}_{BG})_{\epsilon=0} + \frac{1}{2}(\tilde{w}_{GG})_{\epsilon=0}$	$\frac{1}{2}(\tilde{w}_{BG})_{\epsilon=0} + \frac{1}{2}(\tilde{w}_{GG})_{\epsilon=0}$
G	G	G	G	$(\tilde{w}_{GG})_{\epsilon=0}$	$(\tilde{w}_{GG})_{\epsilon=0}$

Each limit is thus a convex combination of $(\tilde{w}_{BG})_{\epsilon=0}$ and $(\tilde{w}_{GG})_{\epsilon=0}$, with a weight h_0 that depends on the mutant bits ($\tilde{d}_{BB}, \tilde{d}_{BG}, \tilde{d}_{GB}, \tilde{d}_{GG}$) and, through $g(\chi)$, on the resident class. Since $\tilde{p} = \text{ALLD}$ and $p_{GG} = C$ by Eq. (96), the two payoffs are

$$(\tilde{w}_{BG})_{\epsilon=0} = v(\tilde{p}_{BG}, p_{GB}) = v(D, p_{GB}) = \begin{cases} v_{DD} = P & \text{if } p_{GB} = D \\ v_{DC} = T & \text{if } p_{GB} = C \end{cases}, \quad (100a)$$

$$(\tilde{w}_{GG})_{\epsilon=0} = v(\tilde{p}_{GG}, p_{GG}) = v_{DC} = T. \quad (100b)$$

Table 4 reports $\lim_{\chi \rightarrow 0} \tilde{w}(\chi)$ for every combination of the mutant assessment bits, separately for the two mutually exclusive resident classes $(p, d) \in \mathcal{A}_{G^{**}G}$ and $(p, d) \in \mathcal{A}_{BGGG} \cup \mathcal{A}_{GGGG}$.

If $p_{GB} = C$, then $(\tilde{w}_{BG})_{\epsilon=0} = (\tilde{w}_{GG})_{\epsilon=0} = T$ by Eq. (100); since $\lim_{\chi \rightarrow 0} \tilde{w}(\chi)$ is a convex combination of these two payoffs, it equals T regardless of the mutant bits and the resident class (Table 4). As $T > R$ for prisoner's dilemmas and snowdrift games, ALLD invades there, and a universally stable righteous system must therefore satisfy

$$p_{GB} = D. \quad (101)$$

Under condition (101), Eq. (100) gives $(\tilde{w}_{BG})_{\epsilon=0} = P$ and $(\tilde{w}_{GG})_{\epsilon=0} = T$; the limit $\lim_{\chi \rightarrow 0} \tilde{w}(\chi) = (1 - h_0)P + h_0T$ is then a convex combination of P and T . Since $R > P$ (3) always holds and $T > R$ holds for prisoner's dilemmas and snowdrift games (so $P < R < T$ there), Table 4 leaves three possibilities:

1. $\lim_{\chi \rightarrow 0} \tilde{w}(\chi) = T$: since $T > R$, ALLD invades for prisoner's dilemmas and snowdrift games.
2. $\lim_{\chi \rightarrow 0} \tilde{w}(\chi)$ lies strictly between P and T : it then exceeds R for some prisoner's dilemmas and snowdrift games, so ALLD can invade.
3. $\lim_{\chi \rightarrow 0} \tilde{w}(\chi) = P$: since $R > P$, the resident resists invasion by ALLD across all cooperative dilemmas.

We conclude that a necessary and sufficient condition for $(p, d) \in \mathcal{R}$ to resist invasion attempts by ALLD across cooperative dilemmas is

$$\lim_{\chi \rightarrow 0} \tilde{w}(\chi) = (\tilde{\omega}_{BG})_{\epsilon=0} = P. \quad (102)$$

Equivalently, $h_0 = 0$: the invading ALLD mutants acquire a bad reputation in the limit, so residents defect against them and they earn only $(\tilde{\omega}_{BG})_{\epsilon=0} = P < R$. We now translate this condition into constraints on the bits of the resident action rule p and assessment rule d .

We distinguish two mutually exclusive cases: (i) $(p, d) \in \mathcal{A}_{BGGG} \cup \mathcal{A}_{GGGG}$, and (ii) $(p, d) \in \mathcal{A}_{G^{**}G}$. In the latter, we also use the identity

$$\tilde{d}_{GB} = d(G, B, \tilde{p}_{GB}, p_{BG}) = d(G, B, p_{GB}, p_{BG}) = d_{GB}, \quad (103)$$

which follows from Eqs. (101) and (98).

1. Suppose that $(p, d) \in \mathcal{A}_{BGGG} \cup \mathcal{A}_{GGGG}$. Inspection of Table 4 shows that condition (102) holds if and only if

$$\tilde{d}_{BG} = d(B, G, \tilde{p}_{BG}, p_{GB}) = d_{BGDD} = B \quad (104)$$

and

$$\tilde{d}_{GG} = d(G, G, \tilde{p}_{GG}, p_{GG}) = d_{GGDC} = B \quad (105)$$

both hold.

2. Suppose that $(p, d) \in \mathcal{A}_{G^{**}G}$, so that $d_{BB} = d_{GG} = G$ and $d_{BG} \neq d_{GB}$ (Eq. (70)). By the identity (103), $\tilde{d}_{GB} = d_{GB}$, giving two subcases.

- (a) If $(d_{BG}, d_{GB}) = (B, G)$, then $\tilde{d}_{GB} = d_{GB} = G$, and Table 4 shows that condition (102) holds if and only if both (104) and (105) hold — the same condition as in case (i).
- (b) If $(d_{BG}, d_{GB}) = (G, B)$, then $\tilde{d}_{GB} = d_{GB} = B$, and Table 4 shows that condition (102) holds if and only if (104) holds, together with either (105) or

$$\tilde{d}_{BB} = d(B, B, \tilde{p}_{BB}, p_{BB}) = d(B, B, D, p_{BB}) = B.$$

Together with the necessary condition $p_{GB} = D$ (101), these cases give condition (ii) of Proposition 11 when $d_{GB} = G$ (case (i) and subcase (ii-a)) and of Proposition 12 when $d_{GB} = B$ (subcase (ii-b)). $\square$

These propositions have a further consequence for righteous systems with $d_{BG} = G$, which we use repeatedly in the analyses that follow.

Corollary 4 (Stability against ALLD forces $p_{BG} = C$ when $d_{BG} = G$). *Let $\epsilon = 0$ and $\chi \rightarrow 0$. If a righteous system (p, d) with $d_{BG} = G$ resists invasion attempts by ALLD across cooperative dilemmas, then $p_{BG} = C$.*

Proof. Suppose instead that $p_{BG} = D$. Resistance to ALLD forces $p_{GB} = D$ (101), so that $d_{BG} = d(B, G, p_{BG}, p_{GB}) = d(B, G, D, D) = d_{BGDD}$; with $d_{BG} = G$ this gives $d_{BGDD} = G$, contradicting the condition $d_{BGDD} = B$ that resistance requires (Propositions 11 and 12). $\square$

We denote by $\mathcal{R}'$ the subset of righteous systems that resist invasion by ALLD across cooperative dilemmas, as characterized by Propositions 11 and 12. Combining these two propositions with Eq. (80):

$$\begin{aligned} \mathcal{R}' = \{ & (p, d) \in \mathcal{R} \mid p_{GB} = D, d_{GB} = G, d_{BGDD} = B, d_{GGDC} = B \} \\ & \cup \{ (p, d) \in \mathcal{R} \mid p_{GB} = D, d_{GB} = B, d_{BGDD} = B, \\ & [d_{GGDC} = B \text{ or } d(B, B, D, p_{BB}) = B] \}. \end{aligned} \quad (106)$$

Since systems in $\mathcal{R}'$ satisfy both $p_{GG} = C$ and $p_{GB} = D$, every $(p, d) \in \mathcal{R}'$ has a discriminating action rule. This characterization lets us immediately locate two families of systems of special interest within the subset $\mathcal{R}'$. First, we can verify that all leading-eight systems satisfy the conditions defining $\mathcal{R}'$. Second, among the four subsets of consistent discriminating systems, two belong to $\mathcal{R}'$ whereas the other two are excluded already at this stage.

Proposition 13 (Leading-eight systems belong to $\mathcal{R}'$). $\mathcal{L} \subset \mathcal{R}'$.

Proof. All leading-eight action rules are discriminating, so $p_{GB} = D$, and their assessment rules are third-order with $d_{BGD*} = B$, $d_{GGD*} = B$, and $d_{GBD*} = G$ (Eqs. (49)–(57)). Since the norm ignores the co-player's action, $d_{BGDD} = B$, $d_{GGDC} = B$, and $d_{GB} = d(G, B, D, p_{BG}) = d_{GBD*} = G$. Together with $\mathcal{L} \subset \mathcal{R}$ (Proposition 10), these satisfy the $d_{GB} = G$ branch of Eq. (106). $\square$

Proposition 14 (Consistent discriminating systems and $\mathcal{R}'$). $\mathcal{Q}_{CDISC} \cup \mathcal{Q}_{DISC} \subset \mathcal{R}'$, while $\mathcal{Q}_{PDISC} \cap \mathcal{R}' = \mathcal{Q}_{ODISC} \cap \mathcal{R}' = \emptyset$.

Proof. For $\mathcal{Q}_{CDISC}$ and $\mathcal{Q}_{DISC}$: both are discriminating with $p_{BG} = C$, so that $d_{GB} = d(G, B, D, C) = d_{GBDC}$; Eqs. (45) and (44) give $d_{BGDD} = B$, $d_{GGDC} = B$, and $d_{GBDC} = G$, whence $d_{GB} = G$ and the $d_{GB} = G$ branch of Eq. (106) is satisfied. For $\mathcal{Q}_{PDISC}$ and $\mathcal{Q}_{ODISC}$: Eqs. (47) and (46) include $d_{BGDD} = G$, which violates the necessary condition $d_{BGDD} = B$ for $\mathcal{R}'$ membership. $\square$

The characterization (106) also determines the cardinality of $\mathcal{R}'$.

Proposition 15 (Cardinality of $\mathcal{R}'$). The set $\mathcal{R}'$ contains $|\mathcal{R}'| = 12,800$ moral systems.

Proof. We count the admissible systems as in the proof of Proposition 5, recalling that a condition leaving k bits of d unconstrained is satisfied by 2^k assessment rules. By Eq. (106) every system in $\mathcal{R}'$ has $p_{GG} = C$ and $p_{GB} = D$, leaving four action rules indexed by $(p_{BB}, p_{BG}) \in \{C, D\}^2$; Eq. (106) is moreover the disjoint union of a branch with $d_{GB} = G$ and one with $d_{GB} = B$. For each action rule and branch, the defining conditions fix or restrict a few bits of d and leave the rest free. For instance, in the branch $d_{GB} = G$ with $p_{BG} = C$, the bits $d_{GB} = d(G, B, D, C)$, $d_{BG} = d(B, G, C, D)$ and $d_{BGDD} = d(B, G, D, D)$ are distinct; the conditions $d_{GG} = G$, $d_{GGDC} = B$, $d_{GB} = G$, $d_{BGDD} = B$ fix four bits, and $d_{BB} + d_{BG} + d_{GB} \geq 2$ (with $d_{GB} = G$) admits three of the four assignments of (d_{BB}, d_{BG}) , giving $3 \cdot 2^{10} = 3,072$. Coincidences among the constrained bits occur when $p_{BG} = D$ (then $d_{BG} = d_{BGDD}$) or $p_{BB} = D$ (then $d_{BB} = d(B, B, D, p_{BB})$) and lower the counts; in particular, the $d_{GB} = B$ branch is empty when $p_{BG} = D$, since then $d_{BG} = d_{BGDD} = B$ and $d_{GB} = B$ would violate $d_{BB} + d_{BG} + d_{GB} \geq 2$. Table 5 collects all cases; they sum to $|\mathcal{R}'| = 12,800$. $\square$

Table 5: Number of assessment rules d yielding a system in $\mathcal{R}'$, for each of the four action rules (p_{BB}, p_{BG}) compatible with $p_{GG} = C$, $p_{GB} = D$, split by the two branches ($d_{GB} = G$ and $d_{GB} = B$) of Eq. (106).

p_{BB}	p_{BG}	$d_{GB} = G$	$d_{GB} = B$	total
C	C	3,072	1,536	4,608
C	D	2,048	0	2,048
D	C	3,072	1,024	4,096
D	D	2,048	0	2,048
$ \mathcal{R}' =$				12,800

3.5.3 Invasion attempts by ALLC

For the mutant $\tilde{p} = \text{ALLC}$,

$$\tilde{p}_{BB} = \tilde{p}_{BG} = \tilde{p}_{GB} = \tilde{p}_{GG} = C. \quad (107)$$

We seek which systems in $\mathcal{R}'$ also resist such invasion across cooperative dilemmas. These are characterized by the following three propositions: the first shows that resistance requires $d_{GB} = G$; the other two then characterize the systems with $d_{GB} = G$, according to whether $d_{BG} = B$ or $d_{BG} = G$. We prove all three together below.

Proposition 16 (Stability against ALLC requires $d_{GB} = G$). *Let $\epsilon = 0$ and $\chi \rightarrow 0$. If a righteous system (p, d) that resists invasion attempts by ALLD also resists invasion attempts by ALLC across cooperative dilemmas, then $d_{GB} = G$.*

Proposition 16 implies that $(p, d) \in \mathcal{R} \cap \mathcal{A}_{GGBG}$ cannot be universally stable.

Proposition 17 (Stability against ALLC for righteous systems with $d_{GB} = G$ and $d_{BG} = B$). *Let $\epsilon = 0$ and $\chi \rightarrow 0$. A righteous system (p, d) with $d_{GB} = G$ and $d_{BG} = B$ that resists invasion attempts by ALLD also resists invasion attempts by ALLC across cooperative dilemmas if and only if*

- (i) $d_{BGCD} = B$, and
- (ii) either $d(G, B, C, p_{BG}) = B$ or $d(B, B, C, p_{BB}) = B$.

Proposition 18 (Stability against ALLC for righteous systems with $d_{GB} = G$ and $d_{BG} = G$). *Let $\epsilon = 0$ and $\chi \rightarrow 0$. A righteous system (p, d) with $d_{GB} = G$ and $d_{BG} = G$ that resists invasion attempts by ALLD also resists invasion attempts by ALLC across cooperative dilemmas if and only if $d_{GBCC} = B$.*

Proof of Propositions 16, 17, and 18. Since the resident is discriminating, the resident payoffs are

$$(\omega_{BG})_{\epsilon=0} = v(p_{BG}, D) = \begin{cases} v_{DD} = P & \text{if } p_{BG} = D \\ v_{DC} = S & \text{if } p_{BG} = C \end{cases}, \quad (108a)$$

$$(\omega_{GB})_{\epsilon=0} = v(D, p_{BG}) = \begin{cases} v_{DD} = P & \text{if } p_{BG} = D \\ v_{DC} = T & \text{if } p_{BG} = C \end{cases}, \quad (108b)$$

$$(\omega_{GG})_{\epsilon=0} = v(p_{GG}, p_{GG}) = v_{CC} = R, \quad (108c)$$

Table 6: Puiseux-series expansion of $\tilde{w}(\chi)$ for invasion attempts by ALLC against resident systems in $(p, d) \in \mathcal{A}_{G^{**}G}$. Rows enumerate all possible combinations of the mutant assessment bits $(\tilde{d}_{BB}, \tilde{d}_{BG}, \tilde{d}_{GB}, \tilde{d}_{GG})$ consistent with $\tilde{d}_{GG} = G$ in Eq. (109). Entries are written using $(\tilde{\omega}_{GG})_{\epsilon=0} = v_{CC} = R$ and $(\tilde{\omega}_{BG})_{\epsilon=0} = v_{CD} = S$.

$\tilde{d}_{BB}$	$\tilde{d}_{BG}$	$\tilde{d}_{GB}$	$\tilde{d}_{GG}$	$\tilde{w}(\chi)$
B	B	B	G	$S + O(\sqrt{\chi})$
B	B	G	G	$(S + R)/2 + O(\sqrt{\chi})$
B	G	B	G	$R + [S + (\tilde{\omega}_{GB})_{\epsilon=0} - 2R] \sqrt{\chi} + O(\chi)$
B	G	G	G	$R + [(\tilde{\omega}_{GB})_{\epsilon=0} - R] \sqrt{\chi} + O(\chi)$
G	B	B	G	$(S + R)/2 + O(\sqrt{\chi})$
G	B	G	G	$R + [S + (\tilde{\omega}_{GB})_{\epsilon=0} - 2R] \sqrt{\chi} + O(\chi)$
G	G	B	G	$R + [S + (\tilde{\omega}_{GB})_{\epsilon=0} - 2R] \sqrt{\chi} + O(\chi)$
G	G	G	G	$R + [(\tilde{\omega}_{GB})_{\epsilon=0} - R] \sqrt{\chi} + O(\chi)$

while the ALLC mutant always cooperates, so the mutant payoffs are

$$\begin{aligned}
(\tilde{\omega}_{BG})_{\epsilon=0} &= v(\tilde{p}_{BG}, p_{GB}) = v_{CD} = S, \\
(\tilde{\omega}_{GB})_{\epsilon=0} &= v(\tilde{p}_{GB}, p_{BG}) = v(C, p_{BG}) = \begin{cases} v_{CD} = S & \text{if } p_{BG} = D \\ v_{CC} = R & \text{if } p_{BG} = C \end{cases}, \\
(\tilde{\omega}_{GG})_{\epsilon=0} &= v(\tilde{p}_{GG}, p_{GG}) = v_{CC} = R.
\end{aligned}$$

In both cases $g(\chi) \rightarrow 1$ as $\chi \rightarrow 0$, so $w(\chi) \rightarrow (\omega_{GG})_{\epsilon=0} = R$; the resident resists ALLC exactly when the leading term of $\Delta w(\chi) = w(\chi) - \tilde{w}(\chi)$ is positive, which we obtain by expanding $w(\chi)$ and $\tilde{w}(\chi)$ as $\chi \rightarrow 0$. Because $g(\chi)$ takes different forms on $\mathcal{A}_{G^{**}G}$ and on $\mathcal{A}_{BGGG} \cup \mathcal{A}_{GGGG}$ (Eq. (90)), we treat cases (i) and (ii) separately. Since $\tilde{p}_{GG} = p_{GG}$ (Eqs. (96) and (107)) and $d_{GG} = G$ (97a), we have

$$\tilde{d}_{GG} = d(G, G, \tilde{p}_{GG}, p_{GG}) = d(G, G, p_{GG}, p_{GG}) = d_{GG} = G \quad (109)$$

in both cases.

1. Suppose that $(p, d) \in \mathcal{A}_{G^{**}G}$. The first-order Puiseux series expansion of $w(\chi)$ (93) is then

$$w(\chi) = R + [(\omega_{BG})_{\epsilon=0} + (\omega_{GB})_{\epsilon=0} - 2R] \sqrt{\chi} + O(\chi). \quad (110)$$

Table 6 reports the corresponding expansion of $\tilde{w}(\chi)$ for each combination of the bits $\tilde{d}_{BB}$, $\tilde{d}_{BG}$, and $\tilde{d}_{GB}$, using $\tilde{d}_{GG} = G$ (109), $(\tilde{\omega}_{GG})_{\epsilon=0} = v_{CC} = R$, and $(\tilde{\omega}_{BG})_{\epsilon=0} = v_{CD} = S$. The entries take three forms: $\tilde{w}(\chi)$ may have a leading term below R (the rows equal to S or $(S + R)/2$), so that the resident resists; or it may match the leading term R of $w(\chi)$, so that the $\sqrt{\chi}$ -coefficients decide (the two remaining forms).

- (a) The expansion of $\tilde{w}(\chi)$ is either $S + O(\sqrt{\chi})$ or $(S + R)/2 + O(\sqrt{\chi})$. In both cases, the leading term of $\tilde{w}(\chi)$ is strictly smaller than R , the leading term of $w(\chi)$ in Eq. (110), because $R > S$ holds across cooperative dilemmas (4). Thus, the resident system (p, d) resists invasion attempts by ALLC.

- (b) The expansion of $\tilde{w}(\chi)$ is given by $R + [S + (\omega_{\text{GB}})_{\epsilon=0} - 2R] \sqrt{\chi} + O(\chi)$. The leading term of $\Delta w(\chi)$ is

$$(\omega_{\text{BG}})_{\epsilon=0} + (\omega_{\text{GB}})_{\epsilon=0} - S - (\tilde{\omega}_{\text{GB}})_{\epsilon=0}. \quad (111)$$

If $\mathbf{p}_{\text{BG}} = \text{D}$, then $(\omega_{\text{BG}})_{\epsilon=0} = (\omega_{\text{GB}})_{\epsilon=0} = P$ and $(\tilde{\omega}_{\text{GB}})_{\epsilon=0} = S$, so Eq. (111) reduces to $2(P - S)$, which is negative for snowdrift games. If $\mathbf{p}_{\text{BG}} = \text{C}$, then $(\omega_{\text{BG}})_{\epsilon=0} = S$, $(\omega_{\text{GB}})_{\epsilon=0} = T$, and $(\tilde{\omega}_{\text{GB}})_{\epsilon=0} = R$, so Eq. (111) reduces to $T - R$, which is negative for stag hunts. In either case, the resident system is not universally stable.

- (c) The expansion of $\tilde{w}(\chi)$ is given by $R + [(\tilde{\omega}_{\text{GB}})_{\epsilon=0} - R] \sqrt{\chi} + O(\chi)$. The leading term of $\Delta w(\chi)$ is

$$(\omega_{\text{BG}})_{\epsilon=0} + (\omega_{\text{GB}})_{\epsilon=0} - R - (\tilde{\omega}_{\text{GB}})_{\epsilon=0}. \quad (112)$$

If $\mathbf{p}_{\text{BG}} = \text{D}$, then $(\omega_{\text{BG}})_{\epsilon=0} = (\omega_{\text{GB}})_{\epsilon=0} = P$ and $(\tilde{\omega}_{\text{GB}})_{\epsilon=0} = S$, so Eq. (112) reduces to $2P - R - S$, which fails to be positive unless $P > (S + R)/2$; this condition is violated for all snowdrift games and for some prisoner's dilemmas and stag hunts. If $\mathbf{p}_{\text{BG}} = \text{C}$, then $(\omega_{\text{BG}})_{\epsilon=0} = S$, $(\omega_{\text{GB}})_{\epsilon=0} = T$, and $(\tilde{\omega}_{\text{GB}})_{\epsilon=0} = R$, so Eq. (112) reduces to $S + T - 2R < 0$ by condition (6), so ALLC always invades. In either case, the resident system is not universally stable.

Within the class $(p, d) \in \mathcal{A}_{\text{G}^{**}\text{G}}$, the only case in which the resident system resists invasion is case 1 above. Inspecting Table 6, this holds if and only if

$$\tilde{\mathbf{d}}_{\text{BG}} = d(\text{B}, \text{G}, \tilde{\mathbf{p}}_{\text{BG}}, \mathbf{p}_{\text{GB}}) = \mathbf{d}_{\text{BGCD}} = \text{B} \quad (113)$$

and either $\tilde{\mathbf{d}}_{\text{GB}} = d(\text{G}, \text{B}, \text{C}, \mathbf{p}_{\text{BG}}) = \text{B}$ or $\tilde{\mathbf{d}}_{\text{BB}} = d(\text{B}, \text{B}, \text{C}, \mathbf{p}_{\text{BB}}) = \text{B}$ also holds. Furthermore, condition (113) forces $\mathbf{d}_{\text{BG}} = \text{B}$: if instead $\mathbf{d}_{\text{BG}} = \text{G}$, then either $\mathbf{p}_{\text{BG}} = \text{C}$ (giving $\tilde{\mathbf{d}}_{\text{BG}} = \mathbf{d}_{\text{BG}} = \text{G}$, contradicting condition (113)) or $\mathbf{p}_{\text{BG}} = \text{D}$ (giving $\mathbf{d}_{\text{BGDD}} = \text{G}$, contradicting the ALLD-stability condition $\mathbf{d}_{\text{BGDD}} = \text{B}$ from Propositions 11 and 12). Hence, within $\mathcal{A}_{\text{G}^{**}\text{G}}$, resistance against ALLC requires $(p, d) \in \mathcal{A}_{\text{GBGG}}$, i.e., $\mathbf{d}_{\text{GB}} = \text{G}$.

2. Suppose that $(p, d) \in \mathcal{A}_{\text{BGGG}} \cup \mathcal{A}_{\text{GGGG}}$. Then $\mathbf{d}_{\text{BG}} = \text{G}$ holds. By Corollary 4, this forces $\mathbf{p}_{\text{BG}} = \text{C}$, and therefore $(\omega_{\text{BG}})_{\epsilon=0} = (\tilde{\omega}_{\text{BG}})_{\epsilon=0} = \mathbf{v}_{\text{CD}} = S$, $(\omega_{\text{GB}})_{\epsilon=0} = \mathbf{v}_{\text{DC}} = T$, and $(\tilde{\omega}_{\text{GB}})_{\epsilon=0} = \mathbf{v}_{\text{CC}} = R$. Using also Eq. (99), the first-order Taylor series expansion of $w(\chi)$ in Eq. (93) reduces to

$$w(\chi) = R + (S + T - 2R)\chi + O(\chi^2).$$

Table 7 reports the corresponding expansion of $\tilde{w}(\chi)$ for each combination of $\tilde{\mathbf{d}}_{\text{BB}}$ and $\tilde{\mathbf{d}}_{\text{GB}}$, using $\tilde{\mathbf{d}}_{\text{GG}} = \text{G}$ (109) and $\tilde{\mathbf{d}}_{\text{BG}} = \mathbf{d}_{\text{BG}} = \text{G}$ (since $\mathbf{p}_{\text{BG}} = \text{C}$).

Both entries of Table 7 match the leading term R of $w(\chi)$, so the χ -coefficients decide; there are two cases.

- (a) The expansion of $\tilde{w}(\chi)$ is given by $R + (S - R)\chi + O(\chi^2)$. In this case, the leading term of $\Delta w(\chi)$ is $T - R$, which is negative for stag hunts. Thus, the resident system can be invaded by ALLC for that class of cooperative dilemmas, and hence is not universally stable.
- (b) The expansion of $\tilde{w}(\chi)$ is given by $R + 2(S - R)\chi + O(\chi^2)$. In this case, the leading term of $\Delta w(\chi)$ is given by $T - S$, which is positive across cooperative dilemmas since $T > S$ always holds. Hence, the resident system (p, d) resists invasion attempts by ALLC.

Table 7: Taylor-series expansion of $\tilde{w}(\chi)$ for invasion attempts by ALLC against resident systems in $(p, d) \in \mathcal{A}_{\text{BGGG}} \cup \mathcal{A}_{\text{GGGG}}$. Rows enumerate all possible combinations of the mutant assessment bits $(\tilde{d}_{\text{BB}}, \tilde{d}_{\text{BG}}, \tilde{d}_{\text{GB}}, \tilde{d}_{\text{GG}})$ consistent with $\tilde{d}_{\text{BG}} = \tilde{d}_{\text{GG}} = \text{G}$.

$\tilde{d}_{\text{BB}}$	$\tilde{d}_{\text{BG}}$	$\tilde{d}_{\text{GB}}$	$\tilde{d}_{\text{GG}}$	$\tilde{w}(\chi)$
B	G	B	G	$R + 2(S - R)\chi + O(\chi^2)$
B	G	G	G	$R + (S - R)\chi + O(\chi^2)$
G	G	B	G	$R + 2(S - R)\chi + O(\chi^2)$
G	G	G	G	$R + (S - R)\chi + O(\chi^2)$

Within the class $(p, d) \in \mathcal{A}_{\text{BGGG}} \cup \mathcal{A}_{\text{GGGG}}$, the resident system resists invasion only in case 2 above. By inspection of Table 7, this holds if and only if

$$\tilde{d}_{\text{GB}} = d(\text{G}, \text{B}, \tilde{p}_{\text{GB}}, p_{\text{BG}}) = d_{\text{GBCC}} = \text{B}.$$

Together, the two cases establish the three propositions: case (i) rules out $\mathcal{A}_{\text{GGBG}}$ — so resistance requires $d_{\text{GB}} = \text{G}$ (Proposition 16) — and characterizes the $d_{\text{BG}} = \text{B}$ systems of $\mathcal{A}_{\text{GBGG}}$ (Proposition 17), while case (ii) characterizes the systems with $d_{\text{GB}} = d_{\text{BG}} = \text{G}$ via $d_{\text{GBCC}} = \text{B}$ (Proposition 18). $\square$

We denote by $\mathcal{R}''$ the subset of $\mathcal{R}'$ that resists invasion by ALLC across cooperative dilemmas, as characterized by Propositions 16, 17 and 18. Proposition 16 forces $d_{\text{GB}} = \text{G}$ for all systems in $\mathcal{R}''$, and the remaining conditions split on d_{BG} :

$$\begin{aligned} \mathcal{R}'' = \{ & (p, d) \in \mathcal{R}' \mid d_{\text{GB}} = \text{G}, d_{\text{BG}} = \text{B}, d_{\text{BGCD}} = \text{B}, \\ & [d(\text{G}, \text{B}, \text{C}, p_{\text{BG}}) = \text{B} \text{ or } d(\text{B}, \text{B}, \text{C}, p_{\text{BB}}) = \text{B}] \} \\ & \cup \{ (p, d) \in \mathcal{R}' \mid d_{\text{GB}} = \text{G}, d_{\text{BG}} = \text{G}, d_{\text{GBCC}} = \text{B} \}. \end{aligned} \quad (114)$$

As for $\mathcal{R}'$, we locate the leading-eight and consistent discriminating systems within $\mathcal{R}''$. Resistance to ALLC now removes half of the leading eight, whereas the consistent discriminating families split exactly as they did for $\mathcal{R}'$.

Proposition 19 (Leading-eight systems and $\mathcal{R}''$). *A leading-eight system belongs to $\mathcal{R}''$ if and only if $d_{\text{GBC}^*} = \text{B}$. The systems satisfying this condition are $\{\text{L2}, \text{L5}, \text{L6}, \text{L8}\}$; the remaining four systems $\{\text{L1}, \text{L3}, \text{L4}, \text{L7}\}$ belong to $\mathcal{R}' \setminus \mathcal{R}''$.*

Proof. By Proposition 13, all leading-eight systems belong to $\mathcal{R}'$ with $d_{\text{GB}} = \text{G}$. Since all have $d_{\text{BGC}^*} = \text{G}$ (Eqs. (49)–(57)), it follows that $d_{\text{BG}} = d_{\text{BGCD}} = \text{G}$, so every leading-eight system falls into the second branch of Eq. (114), which requires $d_{\text{GBCC}} = \text{B}$. Since the norm ignores the co-player's action, $d_{\text{GBCC}} = d_{\text{GBC}^*}$, so the condition $d_{\text{GBCC}} = \text{B}$ is just $d_{\text{GBC}^*} = \text{B}$. Inspecting the definitions: L2, L5, L6, and L8 have $d_{\text{GBC}^*} = \text{B}$; L1, L3, L4, and L7 have $d_{\text{GBC}^*} = \text{G}$. $\square$

Proposition 20 (Consistent discriminating systems and $\mathcal{R}''$). *$\mathcal{Q}_{\text{CDISC}} \cup \mathcal{Q}_{\text{DISC}} \subset \mathcal{R}''$, while $\mathcal{Q}_{\text{PDISC}} \cap \mathcal{R}'' = \mathcal{Q}_{\text{ODISC}} \cap \mathcal{R}'' = \emptyset$.*

Proof. Since $\mathcal{Q}_{\text{PDISC}} \cap \mathcal{R}' = \mathcal{Q}_{\text{ODISC}} \cap \mathcal{R}' = \emptyset$ (Proposition 14) and $\mathcal{R}'' \subset \mathcal{R}'$, the same holds for $\mathcal{R}''$. For $\mathcal{Q}_{\text{CDISC}}$ and $\mathcal{Q}_{\text{DISC}}$: both have $p_{\text{BG}} = \text{C}$, so $d_{\text{BG}} = d(\text{B}, \text{G}, \text{C}, \text{D}) = d_{\text{BGCD}}$ and $d_{\text{GB}} = d(\text{G}, \text{B}, \text{D}, \text{C}) = d_{\text{GBDC}}$; Eqs. (45) and (44) give $d_{\text{BGCD}} = \text{G}$, $d_{\text{GBDC}} = \text{G}$, and $d_{\text{GBCC}} = \text{B}$, whence $d_{\text{BG}} = d_{\text{GB}} = \text{G}$ and the second branch of Eq. (114) is satisfied. $\square$

Table 8: Number of assessment rules d yielding a system in $\mathcal{R}''$, for each of the four action rules (p_{BB}, p_{BG}) , split by the two branches ($d_{BG} = G$ and $d_{BG} = B$) of Eq. (114).

p_{BB}	p_{BG}	$d_{BG} = G$	$d_{BG} = B$	total
C	C	1,024	512	1,536
C	D	0	512	512
D	C	1,024	768	1,792
D	D	0	768	768
$ \mathcal{R}'' =$				4,608

The characterization (114) likewise determines the cardinality of $\mathcal{R}''$.

Proposition 21 (Cardinality of $\mathcal{R}''$). *The set $\mathcal{R}''$ contains $|\mathcal{R}''| = 4,608$ moral systems.*

Proof. We count the admissible systems as before, recalling that a condition leaving k bits of d unconstrained is satisfied by 2^k assessment rules. Both branches of Eq. (114) require $d_{GB} = G$, so $\mathcal{R}''$ lies within the $d_{GB} = G$ branch of $\mathcal{R}'$; on top of the conditions imposed there it adds constraints that split on d_{BG} . For instance, with $p_{BG} = C$ and $d_{BG} = G$: since $p_{BG} = C$ we have $d_{BG} = d(B, G, C, D)$, so $d_{BG} = G$ and $d_{GBCC} = B$ fix two further bits, the majority condition $d_{BB} + d_{BG} + d_{GB} \geq 2$ is then automatic and d_{BB} is free; six bits are fixed and ten free, giving $2^{10} = 1,024$. The $d_{BG} = G$ branch is moreover empty when $p_{BG} = D$, since then $d_{BG} = d_{BGDD}$, which $\mathcal{R}'$ already forces to B. Table 8 collects all cases; they sum to $|\mathcal{R}''| = 4,608$. $\square$

A direct enumeration of all 14×2^{16} conditional systems confirms the cardinalities $|\mathcal{R}|$, $|\mathcal{R}'|$, and $|\mathcal{R}''|$ reported in Propositions 5–21 and in Fig. 4.

3.5.4 Invasion attempts by a discriminating mutant

We now determine which systems in $\mathcal{R}''$ also resist invasion by a discriminating mutant $\tilde{p} \neq p$ across cooperative dilemmas. Such a mutant satisfies

$$\tilde{p}_{GG} = C \quad \text{and} \quad \tilde{p}_{GB} = D. \quad (115)$$

The main results are the following two propositions, which we prove in turn below.

Proposition 22 (Invasion by discriminating mutants for righteous systems with $d_{BG} = B$ or $d_{BB} = B$). *Let $\epsilon = 0$ and $\chi \rightarrow 0$. A righteous system (p, d) with $d_{BG} = B$ or $d_{BB} = B$ that resists invasion attempts by ALLD and ALLC cannot resist invasion attempts by at least one alternative discriminating strategy.*

Propositions 16 and 22 together imply that $(p, d) \in \mathcal{R} \cap (\mathcal{A}_{GGBG} \cup \mathcal{A}_{GBGG} \cup \mathcal{A}_{BGGG})$ cannot be universally stable; among the righteous classes, this leaves only $\mathcal{A}_{GGGG}$, which the next proposition characterizes.

Proposition 23 (Stability against discriminating mutants for righteous systems in $\mathcal{A}_{GGGG}$). *Let $\epsilon = 0$ and $\chi \rightarrow 0$. A righteous system (p, d) with $d_{BB} = d_{BG} = d_{GB} = d_{GG} = G$ that resists invasion attempts by ALLD and ALLC also resists invasion attempts by $\tilde{p} \in \{\text{DISC}, \text{CDISC}\}$ with $\tilde{p} \neq p$ across cooperative dilemmas if and only if $p_{BB} = C$ and $d_{BBDC} = B$.*

Preliminary observations. Since both p and $\tilde{p}$ are discriminating, the following equalities hold:

$$(\tilde{\omega}_{GG})_{\epsilon=0} = (\omega_{GG})_{\epsilon=0} = R, \quad (116a)$$

$$(\tilde{\omega}_{GB})_{\epsilon=0} = (\omega_{GB})_{\epsilon=0} = v(p_{GB}, p_{BG}) = v(D, p_{BG}). \quad (116b)$$

Moreover, $\tilde{p}_{GG} = p_{GG}$ and $\tilde{p}_{GB} = p_{GB}$, so from $d_{GG} = G$ (97a) and $d_{GB} = G$ (Proposition 16) we obtain

$$\tilde{d}_{GG} = d(G, G, \tilde{p}_{GG}, p_{GG}) = d(G, G, p_{GG}, p_{GG}) = d_{GG} = G \quad (117)$$

and

$$\tilde{d}_{GB} = d(G, B, \tilde{p}_{GB}, p_{BG}) = d(G, B, p_{GB}, p_{BG}) = d_{GB} = G. \quad (118)$$

These two equalities, $\tilde{d}_{GG} = \tilde{d}_{GB} = G$, hold throughout the analysis below. The mutant assessment bits $\tilde{d}_{BG}$ and $\tilde{d}_{BB}$ depend on which class of systems is considered, and are determined case by case.

Proof of Proposition 22. We show that for each of the classes $\mathcal{A}_{GBGG}$ and $\mathcal{A}_{BGGG}$ there exists a discriminating mutant that invades for some cooperative dilemma.

Case $\mathcal{A}_{GBGG}$ (i.e., $d_{BG} = B$). The Puiseux-series expansion of $w(\chi)$ (93) is

$$w(\chi) = (\omega_{GG})_{\epsilon=0} + [(\omega_{BG})_{\epsilon=0} + (\omega_{GB})_{\epsilon=0} - 2(\omega_{GG})_{\epsilon=0}] \sqrt{\chi} + O(\chi). \quad (119)$$

By Propositions 11 and 17, $d_{BGDD} = d_{BGCD} = B$ holds. Hence, regardless of $\tilde{p}_{BG}$,

$$\tilde{d}_{BG} = d(B, G, \tilde{p}_{BG}, p_{GB}) = B. \quad (120)$$

Consider the discriminating mutant $\tilde{p}$ with $\tilde{p}_{BG} \neq p_{BG}$ and $\tilde{p}_{BB} = p_{BB}$. Then $\tilde{d}_{BB} = d_{BB} = G$, and the mutant assessment bits are $(\tilde{d}_{BB}, \tilde{d}_{BG}, \tilde{d}_{GB}, \tilde{d}_{GG}) = (G, B, G, G)$, so $\tilde{w}(\chi)$ has the same expansion form as $w(\chi)$. The leading term of $\Delta w(\chi)$ reduces to

$$(\omega_{BG})_{\epsilon=0} - (\tilde{\omega}_{BG})_{\epsilon=0}, \quad (121)$$

which equals $P - S$ if $p_{BG} = D$ (negative for snowdrift games) or $S - P$ if $p_{BG} = C$ (negative for prisoner's dilemmas and stag hunts), by Eq. (108a). Hence $(p, d) \in \mathcal{A}_{GBGG}$ cannot be universally stable.

Case $\mathcal{A}_{BGGG}$ (i.e., $d_{BG} = G, d_{BB} = B$). Since $d_{BG} = G$, Corollary 4 gives $p_{BG} = C$. Consider a discriminating mutant $\tilde{p}$ with $\tilde{p}_{BG} = p_{BG}$ but $\tilde{p}_{BB} \neq p_{BB}$. Then

$$\tilde{d}_{BG} = d(B, G, p_{BG}, p_{GB}) = d_{BG} = G \quad (122)$$

and

$$(\tilde{\omega}_{BG})_{\epsilon=0} = (\omega_{BG})_{\epsilon=0} = v_{CD} = S. \quad (123)$$

The second-order expansion of $w(\chi)$ is

$$\begin{aligned} w(\chi) = & (\omega_{GG})_{\epsilon=0} + [(\omega_{BG})_{\epsilon=0} + (\omega_{GB})_{\epsilon=0} - 2(\omega_{GG})_{\epsilon=0}] \chi \\ & + [(\omega_{BB})_{\epsilon=0} - (\omega_{GG})_{\epsilon=0}] \chi^2 + O(\chi^3). \end{aligned} \quad (124)$$

Since $\tilde{d}_{GG} = \tilde{d}_{GB} = \tilde{d}_{BG} = G$, the mutant assessment bits are $(\tilde{d}_{BB}, G, G, G)$ with $\tilde{d}_{BB} \in \{G, B\}$. Computing $\tilde{w}(\chi)$ from (94) and subtracting, the leading term of $\Delta w(\chi)$ at order χ^2 is reported in Table 9 for each subcase.

When $\tilde{d}_{BB} = G$, the leading term evaluates to $S - T < 0$ (if $p_{BB} = C$) or $P - R < 0$ (if $p_{BB} = D$), both always negative. When $\tilde{d}_{BB} = B$, it evaluates to $R - T$ (if $p_{BB} = C$), which is negative for prisoner's dilemmas and snowdrift games, or to $P - S$ (if $p_{BB} = D$), which is negative for snowdrift games. In every case, the mutant invades for at least one class of cooperative dilemmas. Hence $(p, d) \in \mathcal{A}_{BGGG}$ cannot be universally stable. $\square$

Table 9: Leading term of $\Delta w(\chi)$ for a discriminating mutant with $\tilde{p}_{BB} \neq p_{BB}$ invading a resident system in $\mathcal{A}_{BGGG}$. In each row, the expression is evaluated for $p_{BB} = C$ (so $\tilde{p}_{BB} = D$) and $p_{BB} = D$ (so $\tilde{p}_{BB} = C$).

$\tilde{d}_{BB}$	Leading term	$p_{BB} = C$	$p_{BB} = D$
G	$(\omega_{BB})_{\epsilon=0} - R - (\tilde{\omega}_{BB})_{\epsilon=0} + S$	$S - T$	$P - R$
B	$(\omega_{BB})_{\epsilon=0} - (\tilde{\omega}_{BB})_{\epsilon=0}$	$R - T$	$P - S$

Table 10: Leading term of $\Delta w(\chi)$ for a discriminating mutant with $\tilde{p}_{BB} \neq p_{BB}$ invading a resident system in $\mathcal{A}_{GGGG}$. In each row, the expression is evaluated for $p_{BB} = C$ (so $\tilde{p}_{BB} = D$) and $p_{BB} = D$ (so $\tilde{p}_{BB} = C$).

$\tilde{d}_{BB}$	Leading term	$p_{BB} = C$	$p_{BB} = D$
G	$(\omega_{BB})_{\epsilon=0} - (\tilde{\omega}_{BB})_{\epsilon=0}$	$R - T$	$P - S$
B	$(\omega_{BB})_{\epsilon=0} - (\omega_{BG})_{\epsilon=0} + (\omega_{GG})_{\epsilon=0} - (\tilde{\omega}_{BB})_{\epsilon=0}$	$2R - T - S$	$R + P - 2S$

Proof of Proposition 23. Suppose $(p, d) \in \mathcal{A}_{GGGG}$, i.e., $d_{BB} = d_{BG} = d_{GB} = d_{GG} = G$. As in the $\mathcal{A}_{BGGG}$ case, Corollary 4 gives $p_{BG} = C$, and a discriminating mutant $\tilde{p}$ with $\tilde{p}_{BG} = p_{BG}$ and $\tilde{p}_{BB} \neq p_{BB}$ satisfies Eqs. (122) and (123). The second-order expansion of $w(\chi)$ is

$$w(\chi) = (\omega_{GG})_{\epsilon=0} + [(\omega_{BG})_{\epsilon=0} + (\omega_{GB})_{\epsilon=0} - 2(\omega_{GG})_{\epsilon=0}] \chi + [(\omega_{BB})_{\epsilon=0} - (\omega_{BG})_{\epsilon=0} - (\omega_{GB})_{\epsilon=0} + (\omega_{GG})_{\epsilon=0}] \chi^2 + O(\chi^3). \quad (125)$$

The mutant assessment bits are again $(\tilde{d}_{BB}, G, G, G)$ with $\tilde{d}_{BB} \in \{G, B\}$. Computing $\tilde{w}(\chi)$ and subtracting, we obtain the leading terms of $\Delta w(\chi)$ reported in Table 10.

When $\tilde{d}_{BB} = G$, the leading term $(\omega_{BB})_{\epsilon=0} - (\tilde{\omega}_{BB})_{\epsilon=0}$ evaluates to $R - T$ for $p_{BB} = C$ (so $\tilde{p}_{BB} = D$), negative for prisoner's dilemmas and snowdrift games, or to $P - S$ for $p_{BB} = D$ (so $\tilde{p}_{BB} = C$), negative for snowdrift games. In either case the mutant invades for at least one class, so $\tilde{d}_{BB} = B$ is necessary for universal stability.

When $\tilde{d}_{BB} = B$, the leading term equals $(\omega_{BB})_{\epsilon=0} - (\omega_{BG})_{\epsilon=0} + (\omega_{GG})_{\epsilon=0} - (\tilde{\omega}_{BB})_{\epsilon=0}$.

- If $p_{BB} = D$ (i.e., $p = \text{DISC}$, $\tilde{p} = \text{CDISC}$), this reduces to $R + P - 2S$. While positive for prisoner's dilemmas and stag hunts, this quantity is negative for some snowdrift games, so the resident can be invaded.
- If $p_{BB} = C$ (i.e., $p = \text{CDISC}$, $\tilde{p} = \text{DISC}$), this reduces to $2R - T - S$, which is positive across all cooperative dilemmas by condition (6). Hence the resident resists invasion.

We conclude that universal stability requires $(p, d) \in \mathcal{A}_{GGGG}$, $p_{BB} = C$, and also $\tilde{d}_{BB} = d(B, B, \tilde{p}_{BB}, p_{BB}) = d_{BBDC} = B$. $\square$

We denote by $\mathcal{R}'''$ the subset of $\mathcal{R}''$ that resists invasion by any discriminating strategy across cooperative dilemmas, as characterized by Propositions 22 and 23. These, together with Propositions 11, 16, 18 and Corollary 4, imply that the systems in $\mathcal{R}'''$ satisfy $p = \text{CCDC}$ together with $d_{BBCC} = d_{BGCD} = d_{GBDC} = d_{GGCC} = G$ and $d_{BBDC} = d_{BGDD} = d_{GBCC} = d_{GGDC} = B$; by Eq. (45), this set of righteous systems coincides with the contrite consistent discriminating systems, i.e.,

$$\mathcal{R}''' = \mathcal{Q}_{\text{CDISC}}. \quad (126)$$

Since every discriminating action rule admits $2^8 = 256$ consistent assessment rules (Section 2.3), it follows in particular that $|\mathcal{R}'''| = |\mathcal{Q}_{\text{CDISC}}| = 256$, the last of the cardinalities reported in Fig. 4.

As for $\mathcal{R}'$ and $\mathcal{R}''$, we locate the leading-eight and consistent discriminating systems within $\mathcal{R}'''$. Only one leading-eight system survives, and among the consistent discriminating families only $\mathcal{Q}_{\text{CDISC}}$ meets $\mathcal{R}'''$ — in fact equal to it.

Proposition 24 (Leading-eight systems and $\mathcal{R}'''$). *Among the leading eight, only L2 belongs to $\mathcal{R}'''$; the other seven (L1, L3, L4, L5, L6, L7, L8) do not.*

Proof. Since $\mathcal{R}''' = \mathcal{Q}_{\text{CDISC}}$ (126), membership in $\mathcal{R}'''$ requires $p = \text{CCDC}$. Among the leading eight, only L1 and L2 use $p = \text{CCDC}$; all others use $p = \text{DCDC}$ and are thus excluded. L1 fails because $d_{\text{GBC}*} = G$ (i.e., $d_{\text{GBCC}} = G \neq B$), violating the condition $d_{\text{GBCC}} = B$ in Eq. (45). L2 satisfies all conditions of $\mathcal{Q}_{\text{CDISC}}$ (cf. Eqs. (51) and (45)). $\square$

Proposition 25 (Consistent discriminating systems and $\mathcal{R}'''$). *$\mathcal{Q}_{\text{CDISC}} = \mathcal{R}'''$, while $\mathcal{Q}_{\text{DISC}} \cap \mathcal{R}''' = \mathcal{Q}_{\text{PDISC}} \cap \mathcal{R}''' = \mathcal{Q}_{\text{ODISC}} \cap \mathcal{R}''' = \emptyset$.*

Proof. The equality $\mathcal{Q}_{\text{CDISC}} = \mathcal{R}'''$ is Eq. (126). Since $\mathcal{Q}_{\text{DISC}}$, $\mathcal{Q}_{\text{PDISC}}$, and $\mathcal{Q}_{\text{ODISC}}$ use the action rules DCDC, DDDC, and CDDC respectively — all different from CCDC — each is disjoint from $\mathcal{Q}_{\text{CDISC}} = \mathcal{R}'''$. $\square$

3.5.5 The universal stability theorem

We are now in a position to prove the main result of this section.

Theorem 1 (Universal stability within righteous systems). *Let $\epsilon = 0$ and $\chi \rightarrow 0$. A righteous system is universally stable if and only if it is a contrite consistent discriminating system.*

Proof. Necessity. A universally stable righteous system resists ALLD, ALLC, and every discriminating strategy; by the successive eliminations of Sections 3.5.2, 3.5.3, and 3.5.4, it therefore lies in $\mathcal{R}''' = \mathcal{Q}_{\text{CDISC}}$ (126), i.e., it is contrite consistent discriminating.

Sufficiency. Conversely, let $(p, d) \in \mathcal{Q}_{\text{CDISC}}$, so that $p = \text{CDISC}$. We compute $\Delta w(\chi)$ for each of the 15 mutant strategies $\tilde{p} \neq p$; since all systems in $\mathcal{Q}_{\text{CDISC}}$ share the same reputation equilibria and payoffs at $\epsilon = 0$ (Remark 1), $\Delta w(\chi)$ does not depend on which consistent assessment rule accompanies CDISC, and the results, collected in Table 11, hold for all of $\mathcal{Q}_{\text{CDISC}}$ at once.

A direct inspection of Table 11 reveals three conditions for $(p, d) \in \mathcal{Q}_{\text{CDISC}}$ to resist invasion by a mutant strategy $\tilde{p} \neq p$:

1. If $\tilde{p} = *D**$, then CDISC resists invasion attempts if and only if $R > P$, which always holds across cooperative dilemmas (3).
2. If $\tilde{p} = *CCC$, then CDISC resists invasion attempts if and only if $T > S$, which always holds across cooperative dilemmas (12).
3. Otherwise, CDISC resists invasion attempts if and only if $2R > T + S$, which always holds across cooperative dilemmas (6).

As all three conditions hold across cooperative dilemmas, (p, d) resists every mutant strategy and is therefore universally stable. $\square$

Table 11: Taylor series expansion of the difference between resident and mutant expected payoffs for $p = \text{CCDC}$ (i.e., CDISC) under a consistent assessment rule d for different mutant strategies $\tilde{p} \neq p$. Here, no execution errors are considered ($\epsilon = 0$).

$\tilde{p}$ (bitstring)	$\tilde{p}$ (name)	$\Delta w(\chi)$
DDDD	ALLD	$R - P + O(\chi)$
CDDD		$R - P + O(\chi)$
DCDD		$\frac{1}{2}(2R - T - S) + O(\chi)$
CCDD		$\frac{1}{2}(2R - T - S) + O(\chi)$
DDCD		$R - P + O(\chi)$
CDCD		$R - P + O(\chi)$
DCCD		$\frac{1}{2}(2R - T - S) + O(\chi)$
CCCD		$\frac{1}{2}(2R - T - S) + O(\chi)$
DDDC		$\frac{1}{2}(R - P) + O(\chi)$
CDDC		$\frac{1}{3}(R - P) + O(\chi)$
DCDC	DISC	$(2R - T - S)\chi^2 + O(\chi^3)$
CCDC	CDISC	—
DDCC	ALLC	$\frac{2}{3}(R - P) + O(\chi)$
CDCC		$\frac{1}{2}(R - P) + O(\chi)$
DCCC		$(T - S)\chi + O(\chi^2)$
CCCC		$(T - S)\chi + O(\chi^2)$

3.6 Stability analysis of consistent discriminating systems

Consistent discriminating systems occupy a central role in our analysis: they are righteous (Proposition 7), and they form the dense cluster in the upper-right corner of the stability–cooperation scatter plot shown in Fig. 2a of the main text. Section 3.5 establishes that, among righteous systems, universal stability — resistance to invasion across the full space of cooperative dilemmas — is exclusive to consistent discriminating systems with CDISC as action rule (Theorem 1).

We now ask how consistent discriminating systems with the three other action rules (DISC , ODISC , and PDISC) fare. Universal stability being out of reach for these systems, we determine instead the limiting values of their per-class stability indices σ^{PD} , σ^{SG} , and σ^{SH} , thereby identifying the classes of cooperative dilemma (prisoner’s dilemma, snowdrift game, stag hunt) in which each system remains stable. We proceed as in the previous section, expanding the payoff difference $\Delta w(\chi)$ around $\chi = 0$ for every alternative mutant strategy; since all 256 consistent systems sharing an action rule have identical reputation equilibria and payoffs at $\epsilon = 0$ (Remark 1), these expansions depend only on the action rule, and each result below holds for every consistent assessment rule at once. For ease of comparison, we first restate the result for CDISC (Corollary 5) and then treat each remaining case in turn; Table 15 collects the results.

3.6.1 Consistent discriminating systems with CDISC as action rule

Since systems in $\mathcal{Q}_{\text{CDISC}}$ are righteous (Proposition 7) and are precisely the contrite consistent discriminating systems (Definition 7), Theorem 1 implies that they are universally stable, and Definition 12 immediately gives:

Table 12: Taylor series expansion of the difference between resident and mutant expected payoffs for $p = \text{DCDC}$ (i.e., DISC) under a consistent assessment rule d for different mutant strategies $\tilde{p} \neq p$. Here, no execution errors are considered ($\epsilon = 0$).

$\tilde{p}$ (bitstring)	$\tilde{p}$ (name)	$\Delta w(\chi)$
DDDD	ALLD	$R - P + O(\chi)$
CDDD		$R - P + O(\chi)$
DCDD		$\frac{1}{2}(2R - T - S) + O(\chi)$
CCDD		$\frac{1}{2}(2R - T - S) + O(\chi)$
DDCD		$R - P + O(\chi)$
CDCD		$R - P + O(\chi)$
DCCD		$\frac{1}{2}(2R - T - S) + O(\chi)$
CCCD		$\frac{1}{2}(2R - T - S) + O(\chi)$
DDDC		$\frac{1}{3}(R - P) + O(\chi)$
CDDC		$\frac{1}{2}(R - P) + O(\chi)$
DCDC	DISC	—
CCDC	CDISC	$(R + P - 2S)\chi^2 + O(\chi^3)$
DDCC	ALLC	$\frac{1}{2}(R - P) + O(\chi)$
CDCC		$\frac{2}{3}(R - P) + O(\chi)$
DCCC		$(T - S)\chi + O(\chi^2)$
CCCC		$(T - S)\chi + O(\chi^2)$

Corollary 5 (Stability of consistent discriminating systems with CDISC as action rule). *Systems in $\mathcal{Q}_{\text{CDISC}}$ are universally stable, that is,*

$$\lim_{\chi \rightarrow 0} \left(\sigma_{p,d}^{\text{PD}} \right)_{\epsilon=0} = \lim_{\chi \rightarrow 0} \left(\sigma_{p,d}^{\text{SG}} \right)_{\epsilon=0} = \lim_{\chi \rightarrow 0} \left(\sigma_{p,d}^{\text{SH}} \right)_{\epsilon=0} = 1 \quad \text{for all } (p, d) \in \mathcal{Q}_{\text{CDISC}}.$$

3.6.2 Consistent discriminating systems with DISC as action rule

It is clear from the characterizations of the sets $\mathcal{Q}_{\text{DISC}}$ (44) and $\mathcal{R}''$ (114) that $\mathcal{Q}_{\text{DISC}} \subset \mathcal{R}''$ holds, so that a moral system $(p, d) \in \mathcal{Q}_{\text{DISC}}$ resists invasion attempts by ALLD and ALLC across cooperative dilemmas. It is equally clear, from the analysis in Section 3.5.4, that such a system can be invaded by CDISC for some games in our payoff space. The following proposition completes the analysis by considering all alternative mutant strategies and quantifies the resulting stability indices.

Proposition 26 (Stability of consistent discriminating systems with DISC as action rule). *For $\epsilon = 0$ and in the limit $\chi \rightarrow 0$, a consistent discriminating system with DISC as action rule is stable across prisoner's dilemmas and stag hunts, but it can be invaded by CDISC if the cooperative dilemma is a snowdrift game for which $R + P < 2S$ holds. Thus:*

$$\lim_{\chi \rightarrow 0} \left(\sigma_{p,d}^{\text{PD}} \right)_{\epsilon=0} = 1, \quad \lim_{\chi \rightarrow 0} \left(\sigma_{p,d}^{\text{SG}} \right)_{\epsilon=0} = \frac{2}{3}, \quad \lim_{\chi \rightarrow 0} \left(\sigma_{p,d}^{\text{SH}} \right)_{\epsilon=0} = 1 \quad \text{for all } (p, d) \in \mathcal{Q}_{\text{DISC}}.$$

Proof. We compute $\Delta w(\chi)$ for each of the 15 mutant strategies $\tilde{p} \neq p$; the results are collected in Table 12.

A direct inspection of Table 12 reveals four conditions for $(p, d) \in \mathcal{Q}_{\text{DISC}}$ to resist invasion by a mutant strategy $\tilde{p} \neq p$:

1. If $\tilde{p} = *D**$, then DISC resists invasion attempts if and only if $R > P$, which always holds across cooperative dilemmas (3).
2. If $\tilde{p} = *CCC$, then DISC resists invasion attempts if and only if $T > S$, which always holds across cooperative dilemmas (12).
3. If $\tilde{p} = CCDC$ (i.e., CDISC), then DISC resists invasion attempts if and only if $R + P > 2S$, which holds for all prisoner's dilemmas (9) and stag hunts (11), but not for all snowdrift games (10).
4. For all remaining mutant strategies, DISC resists invasion attempts if and only if $2R > T + S$, which always holds across cooperative dilemmas (6).

The conditions in items 1, 2, and 4 hold across all cooperative dilemmas, and the condition in item 3 holds across all prisoner's dilemmas and stag hunts; the stability indices for these two classes thus converge to one. Within snowdrift games, the only binding condition is $R + P > 2S$: it fails precisely for the games with $R + P < 2S$, in which CDISC invades, and by Lemma 1 the snowdrift games satisfying $R + P > 2S$ make up a fraction $2/3$ of the class. $\square$

3.6.3 Consistent discriminating systems with ODISC as action rule

It is clear from the characterizations of the sets Q_{ODISC} (46) and $\mathcal{R}'$ (106) that $Q_{\text{ODISC}} \not\subset \mathcal{R}'$ holds, so that a moral system $(p, d) \in Q_{\text{ODISC}}$ can be invaded by ALLD for some cooperative dilemmas. The following proposition completes the analysis by considering all alternative mutant strategies and quantifies the resulting stability indices.

Proposition 27 (Stability of consistent discriminating systems with ODISC as action rule). *For $\epsilon = 0$ and in the limit $\chi \rightarrow 0$, a consistent discriminating system with ODISC as action rule is stable across snowdrift games and stag hunts, but it can be invaded by mutants with $\tilde{p}_{BG} = \tilde{p}_{GG} = D$, or by PDISC, if the cooperative dilemma is a prisoner's dilemma for which $2R < T + P$ holds. Thus:*

$$\lim_{\chi \rightarrow 0} \left(\sigma_{p,d}^{\text{PD}} \right)_{\epsilon=0} = \frac{2}{3}, \quad \lim_{\chi \rightarrow 0} \left(\sigma_{p,d}^{\text{SG}} \right)_{\epsilon=0} = 1, \quad \lim_{\chi \rightarrow 0} \left(\sigma_{p,d}^{\text{SH}} \right)_{\epsilon=0} = 1 \quad \text{for all } (p, d) \in Q_{\text{ODISC}}.$$

Proof. We compute $\Delta w(\chi)$ for each of the 15 mutant strategies $\tilde{p} \neq p$; the results are collected in Table 13.

A direct inspection of Table 13 reveals two conditions for $(p, d) \in Q_{\text{ODISC}}$ to resist invasion by a mutant strategy $\tilde{p} \neq p$:

1. If $\tilde{p} = *D*D$ or $\tilde{p} = DDDC$ (i.e., PDISC), then ODISC resists invasion attempts if and only if

$$2R > T + P \tag{127}$$

holds. This condition holds for all snowdrift games (10) and stag hunts (11), but not for all prisoner's dilemmas (9).

2. For all remaining mutant strategies, ODISC resists invasion attempts if and only if $R > S$, which always holds across cooperative dilemmas (4).

The condition in item 2 holds across all cooperative dilemmas, and the condition in item 1 holds across all snowdrift games and stag hunts; the stability indices for these two classes thus converge to one. Within prisoner's dilemmas, the only binding condition is $2R > T + P$ (127): it fails precisely for the games with $2R < T + P$, in which the mutants of item 1 invade, and by Lemma 2 the prisoner's dilemmas satisfying $2R > T + P$ make up a fraction $2/3$ of the class. $\square$

Table 13: Taylor series expansion of the difference between resident and mutant expected payoffs for $p = \text{CDDC}$ (i.e., ODISC) under a consistent assessment rule d for different mutant strategies $\tilde{p} \neq p$. Here, no execution errors are considered ($\epsilon = 0$).

$\tilde{p}$ (bitstring)	$\tilde{p}$ (name)	$\Delta w(\chi)$
DDDD	ALLD	$\frac{1}{2}(2R - T - P) + O(\chi)$
CDDD		$\frac{1}{2}(2R - T - P) + O(\chi)$
DCDD		$R - S + O(\chi)$
CCDD		$R - S + O(\chi)$
DDCD		$\frac{1}{2}(2R - T - P) + O(\chi)$
CDCD		$\frac{1}{2}(2R - T - P) + O(\chi)$
DCCD		$R - S + O(\chi)$
CCCD		$R - S + O(\chi)$
DDDC		$(2R - T - P)\chi^2 + O(\chi^3)$
CDDC		—
DCDC	DISC	$\frac{1}{2}(R - S) + O(\chi)$
CCDC	CDISC	$\frac{1}{3}(R - S) + O(\chi)$
DDCC		$(R - S)\chi + O(\chi^2)$
CDCC	ALLC	$(R - S)\chi + O(\chi^2)$
DCCC		$\frac{2}{3}(R - S) + O(\chi)$
CCCC		$\frac{1}{2}(R - S) + O(\chi)$

3.6.4 Consistent discriminating systems with PDISC as action rule

It is clear from the characterizations of the sets $\mathcal{Q}_{\text{PDISC}}$ (47) and $\mathcal{R}'$ (106) that $\mathcal{Q}_{\text{PDISC}} \not\subseteq \mathcal{R}'$ holds, so that a moral system $(p, d) \in \mathcal{Q}_{\text{PDISC}}$ can be invaded by ALLD for some cooperative dilemmas. The following proposition completes the analysis by considering all alternative mutant strategies and quantifies the resulting stability indices.

Proposition 28 (Stability of consistent discriminating systems with PDISC as action rule). *For $\epsilon = 0$ and in the limit $\chi \rightarrow 0$, a consistent discriminating system with PDISC as action rule is stable across snowdrift games and stag hunts, but it can be invaded by mutants with $\tilde{p}_{\text{BG}} = \tilde{p}_{\text{GG}} = D$ if the cooperative dilemma is a prisoner's dilemma for which $2R < T + P$ holds. Thus:*

$$\lim_{\chi \rightarrow 0} \left(\sigma_{p,d}^{\text{PD}} \right)_{\epsilon=0} = \frac{2}{3}, \quad \lim_{\chi \rightarrow 0} \left(\sigma_{p,d}^{\text{SG}} \right)_{\epsilon=0} = 1, \quad \lim_{\chi \rightarrow 0} \left(\sigma_{p,d}^{\text{SH}} \right)_{\epsilon=0} = 1 \quad \text{for all } (p, d) \in \mathcal{Q}_{\text{PDISC}}.$$

Proof. We compute $\Delta w(\chi)$ for each of the 15 mutant strategies $\tilde{p} \neq p$; the results are collected in Table 14.

A direct inspection of Table 14 reveals two conditions for $(p, d) \in \mathcal{Q}_{\text{PDISC}}$ to resist invasion by a mutant strategy $\tilde{p} \neq p$:

1. If $\tilde{p} = *D*D$, then PDISC resists invasion attempts if and only if $2R > T + P$ (127) holds. This condition holds for all snowdrift games (10) and stag hunts (11), but not for all prisoner's dilemmas (9).
2. For all remaining mutant strategies, PDISC resists invasion attempts if and only if $R > S$, which always holds across cooperative dilemmas (4).

Table 14: Taylor series expansion of the difference between resident and mutant expected payoffs for $p = \text{DDDC}$ (i.e., PDISC) under a consistent assessment rule d for different mutant strategies $\tilde{p} \neq p$. Here, no execution errors are considered ($\epsilon = 0$).

$\tilde{p}$ (bitstring)	$\tilde{p}$ (name)	$\Delta w(\chi)$
DDDD	ALLD	$\frac{1}{2}(2R - T - P) + O(\chi)$
CDDD		$\frac{1}{2}(2R - T - P) + O(\chi)$
DCDD		$R - S + O(\chi)$
CCDD		$R - S + O(\chi)$
DDCD		$\frac{1}{2}(2R - T - P) + O(\chi)$
CDCD		$\frac{1}{2}(2R - T - P) + O(\chi)$
DCCD		$R - S + O(\chi)$
CCCD		$R - S + O(\chi)$
DDDC		—
CDDC		$(R - S)\chi^2 + O(\chi^3)$
DCDC	DISC	$\frac{1}{3}(R - S) + O(\chi)$
CCDC	CDISC	$\frac{1}{2}(R - S) + O(\chi)$
DDCC		$(R - S)\chi + O(\chi^2)$
CDCC		$(R - S)\chi + O(\chi^2)$
DCCC		$\frac{1}{2}(R - S) + O(\chi)$
CCCC	ALLC	$\frac{2}{3}(R - S) + O(\chi)$

The condition in item 2 holds across all cooperative dilemmas, and the condition in item 1 holds across all snowdrift games and stag hunts; the stability indices for these two classes thus converge to one. Within prisoner's dilemmas, the only binding condition is $2R > T + P$ (127): it fails precisely for the games with $2R < T + P$, in which the mutants of item 1 invade, and by Lemma 2 the prisoner's dilemmas satisfying $2R > T + P$ make up a fraction $2/3$ of the class. $\square$

3.6.5 Summary: consistent discriminating systems

Table 15 collects the stability results for the four discriminating action rules under a consistent assessment rule (Corollary 5 and Propositions 26, 27, and 28). The results confirm that, among consistent discriminating systems, universal stability is exclusive to those with CDISC as action rule: every other action rule fails in at least one class of cooperative dilemma. Systems with PDISC or ODISC as action rule can be invaded in the strict subset of prisoner's dilemmas for which $2R < T + P$ holds, while systems with DISC as action rule can be invaded by CDISC in the strict subset of snowdrift games for which $R + P < 2S$ holds.

3.7 Stability analysis of the leading eight

We now determine the limiting per-class stability indices of the leading-eight moral systems (Section 2.4). As we shall see, L2 is the only leading-eight system that achieves universal stability; every other member fails in at least one class of cooperative dilemma.

By Propositions 19 and 24, the leading eight are distributed across the hierarchy $\mathcal{R} \supset \mathcal{R}' \supset \mathcal{R}'' \supset \mathcal{R}'''$ of Section 3.5 as follows: L2 belongs to $\mathcal{R}'''$; L5, L6, and L8 belong to $\mathcal{R}'' \setminus \mathcal{R}'''$; and L1, L3, L4, and L7 belong to $\mathcal{R}' \setminus \mathcal{R}''$. This placement organizes the analysis into three cases:

Table 15: Stability of consistent discriminating systems across cooperative dilemmas, for $\epsilon = 0$ and in the limit $\chi \rightarrow 0$. Each entry gives the limiting value of the stability index for the indicated game class: a value of 1 indicates stability across the entire class, and a value of 2/3 indicates stability across a strict subset. The last column gives the average stability index σ (32). Action rules are ordered as in Fig. 4a of the main text.

Action rule	σ^{PD}	σ^{SG}	σ^{SH}	σ
PDISC	2/3	1	1	8/9
ODISC	2/3	1	1	8/9
DISC	1	2/3	1	8/9
CDISC	1	1	1	1

1. L2 and L6. Since L2 belongs to $\mathcal{Q}_{\text{CDISC}}$ and L6 to $\mathcal{Q}_{\text{DISC}}$, their stability indices follow directly from Corollary 5 and Proposition 26.
2. L5 and L8. These systems resist invasion by ALLD and ALLC across cooperative dilemmas, but can be invaded by some discriminating strategy for some games in our payoff space.
3. L1, L3, L4, and L7. These systems resist invasion by ALLD across cooperative dilemmas, but can be invaded by ALLC for some games in our payoff space.

For the six systems in the last two cases, we proceed as in the previous section, expanding the payoff difference $\Delta w(\chi)$ around $\chi = 0$ for every alternative mutant strategy. We treat the eight systems in turn; Table 22 collects the results.

3.7.1 L1 (CDISC, Standing)

L1 belongs to $\mathcal{R}' \setminus \mathcal{R}''$ (Proposition 19); its vulnerability stems from the bit $d_{\text{GBC}^*} = \text{G}$ — the only bit in which its assessment rule differs from L2’s — which exposes it to invasion by *CCC mutants (among them ALLC) under stag-hunt payoffs. The following proposition quantifies the resulting stability indices.

Proposition 29 (Stability of L1). *For $\epsilon = 0$ and in the limit $\chi \rightarrow 0$, L1 is stable across prisoner’s dilemmas and snowdrift games, but it can always be invaded by $\tilde{p} = \text{*CCC}$ if the cooperative dilemma is a stag hunt. Thus:*

$$\lim_{\chi \rightarrow 0} \left(\sigma_{\text{L1}}^{\text{PD}} \right)_{\epsilon=0} = 1, \quad \lim_{\chi \rightarrow 0} \left(\sigma_{\text{L1}}^{\text{SG}} \right)_{\epsilon=0} = 1, \quad \lim_{\chi \rightarrow 0} \left(\sigma_{\text{L1}}^{\text{SH}} \right)_{\epsilon=0} = 0.$$

Proof. We compute $\Delta w(\chi)$ for each of the 15 mutant strategies $\tilde{p} \neq p$; the results are collected in Table 16.

A direct inspection of Table 16 reveals three conditions for L1 to resist invasion by a mutant strategy $\tilde{p} \neq p$:

1. If $\tilde{p} = \text{*D**}$, then L1 resists invasion attempts if and only if $R > P$, which always holds across cooperative dilemmas (3).
2. If $\tilde{p} = \text{*CCC}$, then L1 resists invasion attempts if and only if $T > R$, which holds for all prisoner’s dilemmas (9) and snowdrift games (10), but for no stag hunt (11).

Table 16: Taylor series expansion of the difference between resident and mutant expected payoffs for the leading-eight moral system L1 (CDISC, Standing) for different mutant strategies $\tilde{p} \neq \text{CDISC}$. Here, no execution errors are considered ($\epsilon = 0$).

$\tilde{p}$ (ID)	$\tilde{p}$ (bitstring)	$\tilde{p}$ (name)	$\Delta w(\chi)$
0	DDDD	ALLD	$R - P + O(\chi)$
1	CDDD		$R - P + O(\chi)$
2	DCDD		$\frac{1}{2}(2R - T - S) + O(\chi)$
3	CCDD		$\frac{1}{2}(2R - T - S) + O(\chi)$
4	DDCD		$R - P + O(\chi)$
5	CDCD		$R - P + O(\chi)$
6	DCCD		$\frac{1}{2}(2R - T - S) + O(\chi)$
7	CCCD		$\frac{1}{2}(2R - T - S) + O(\chi)$
8	DDDC	PDISC	$\frac{1}{2}(R - P) + O(\chi)$
9	CDDC	ODISC	$\frac{1}{3}(R - P) + O(\chi)$
10	DCDC	DISC	$(2R - T - S)\chi^2 + O(\chi^3)$
11	CCDC	CDISC	—
12	DDCC		$\frac{1}{2}(R - P) + O(\chi)$
13	CDCC		$\frac{1}{3}(R - P) + O(\chi)$
14	DCCC		$(T - R)\chi + O(\chi^2)$
15	CCCC	ALLC	$(T - R)\chi + O(\chi^2)$

3. For all remaining mutant strategies, L1 resists invasion attempts if and only if $2R > T + S$, which always holds across cooperative dilemmas (6).

The conditions in items 1 and 3 hold across all cooperative dilemmas, and the condition in item 2 holds across all prisoner's dilemmas and snowdrift games; the stability indices for these two classes thus converge to one. In contrast, $T > R$ fails for every stag hunt, in which the mutants of item 2 invade; the stability index for this class thus converges to zero. $\square$

3.7.2 L2 (CDISC, Consistent Standing)

L2 belongs to $\mathcal{Q}_{\text{CDISC}}$ (cf. Eqs. (51) and (45)), the set of contrite consistent discriminating systems. In contrast with L1, its assessment rule judges cooperation with bad co-players as bad ($d_{\text{GBC}^*} = \text{B}$), so a *CCC mutant loses its good standing whenever it cooperates with a bad co-player, and is subsequently exploited by discriminating residents. As a result, the binding stability condition against these mutants is $T > S$ (Table 11) rather than L1's $T > R$ (Table 16); since $T > S$ holds across all cooperative dilemmas (12), L2 resists these mutants even in stag hunts, where L1 is invaded.

Proposition 30 (Stability of L2). *L2 is universally stable, that is,*

$$\lim_{\chi \rightarrow 0} \left(\sigma_{\text{L2}}^{\text{PD}} \right)_{\epsilon=0} = 1, \quad \lim_{\chi \rightarrow 0} \left(\sigma_{\text{L2}}^{\text{SG}} \right)_{\epsilon=0} = 1, \quad \lim_{\chi \rightarrow 0} \left(\sigma_{\text{L2}}^{\text{SH}} \right)_{\epsilon=0} = 1.$$

Proof. Immediate from Corollary 5, since L2 belongs to $\mathcal{Q}_{\text{CDISC}}$. $\square$

Table 17: Taylor series expansion of the difference between resident and mutant expected payoffs for the leading-eight moral system L3 (DISC, Simple Standing) for different mutant strategies $\tilde{p} \neq \text{DISC}$. Here, no execution errors are considered ($\epsilon = 0$).

$\tilde{p}$ (ID)	$\tilde{p}$ (bitstring)	$\tilde{p}$ (name)	$\Delta w(\chi)$
0	DDDD	ALLD	$R - P + O(\chi)$
1	CDDD		$R - P + O(\chi)$
2	DCDD		$\frac{1}{2}(2R - T - S) + O(\chi)$
3	CCDD		$\frac{1}{2}(2R - T - S) + O(\chi)$
4	DDCD		$R - P + O(\chi)$
5	CDCD		$R - P + O(\chi)$
6	DCCD		$\frac{1}{2}(2R - T - S) + O(\chi)$
7	CCCD		$\frac{1}{2}(2R - T - S) + O(\chi)$
8	DDDC	PDISC	$\frac{1}{3}(R - P) + O(\chi)$
9	CDDC	ODISC	$\frac{1}{3}(R - P) + O(\chi)$
10	DCDC	DISC	—
11	CCDC	CDISC	$(P - S)\chi^2 + O(\chi^3)$
12	DDCC		$\frac{1}{3}(R - P) + O(\chi)$
13	CDCC		$\frac{1}{3}(R - P) + O(\chi)$
14	DCCC		$(T - R)\chi + O(\chi^2)$
15	CCCC	ALLC	$(T - R)\chi + O(\chi^2)$

3.7.3 L3 (DISC, Simple Standing)

L3 belongs to $\mathcal{R}' \setminus \mathcal{R}''$ (Proposition 19); its Simple Standing assessment rule has both $\mathbf{d}_{\text{GBC}*} = \mathbf{G}$ and $\mathbf{d}_{\text{BBC}*} = \mathbf{G}$. The former exposes it, like L1, to invasion by *CCC mutants (among them ALLC) under stag-hunt payoffs; the latter exposes it to invasion by CDISC under snowdrift payoffs. The following proposition quantifies the resulting stability indices.

Proposition 31 (Stability of L3). *For $\epsilon = 0$ and in the limit $\chi \rightarrow 0$, L3 is stable across prisoner's dilemmas, but it can always be invaded by CDISC if the cooperative dilemma is a snowdrift game, and by $\tilde{p} = \text{*CCC}$ if the cooperative dilemma is a stag hunt. Thus:*

$$\lim_{\chi \rightarrow 0} \left(\sigma_{\text{L3}}^{\text{PD}} \right)_{\epsilon=0} = 1, \quad \lim_{\chi \rightarrow 0} \left(\sigma_{\text{L3}}^{\text{SG}} \right)_{\epsilon=0} = 0, \quad \lim_{\chi \rightarrow 0} \left(\sigma_{\text{L3}}^{\text{SH}} \right)_{\epsilon=0} = 0.$$

Proof. We compute $\Delta w(\chi)$ for each of the 15 mutant strategies $\tilde{p} \neq p$; the results are collected in Table 17.

A direct inspection of Table 17 reveals four conditions for L3 to resist invasion by a mutant strategy $\tilde{p} \neq p$:

1. If $\tilde{p} = \text{*D**}$, then L3 resists invasion attempts if and only if $R > P$, which always holds across cooperative dilemmas (3).
2. If $\tilde{p} = \text{*CCC}$, then L3 resists invasion attempts if and only if $T > R$, which holds for all prisoner's dilemmas (9) and snowdrift games (10), but for no stag hunt (11).
3. If $\tilde{p} = \text{CCDC}$ (i.e., CDISC), then L3 resists invasion attempts if and only if $P > S$, which holds for all prisoner's dilemmas (9) and stag hunts (11), but for no snowdrift game (10).

Table 18: Taylor series expansion of the difference between resident and mutant expected payoffs for the leading-eight moral system L4 (DISC, SS-SJ) for different mutant strategies $\tilde{p} \neq \text{DISC}$. Here, no execution errors are considered ($\epsilon = 0$).

$\tilde{p}$ (ID)	$\tilde{p}$ (bitstring)	$\tilde{p}$ (name)	$\Delta w(\chi)$
0	DDDD	ALLD	$R - P + O(\chi)$
1	CDDD		$R - P + O(\chi)$
2	DCDD		$\frac{1}{2}(2R - T - S) + O(\chi)$
3	CCDD		$\frac{1}{2}(2R - T - S) + O(\chi)$
4	DDCD		$R - P + O(\chi)$
5	CDCD		$R - P + O(\chi)$
6	DCCD		$\frac{1}{2}(2R - T - S) + O(\chi)$
7	CCCD		$\frac{1}{2}(2R - T - S) + O(\chi)$
8	DDDC	PDISC	$\frac{1}{3}(R - P) + O(\chi)$
9	CDDC	ODISC	$\frac{1}{2}(R - P) + O(\chi)$
10	DCDC	DISC	—
11	CCDC	CDISC	$(R + P - 2S)\chi^2 + O(\chi^3)$
12	DDCC		$\frac{1}{3}(R - P) + O(\chi)$
13	CDCC		$\frac{1}{2}(R - P) + O(\chi)$
14	DCCC		$(T - R)\chi + O(\chi^2)$
15	CCCC	ALLC	$(T - R)\chi + O(\chi^2)$

4. For all remaining mutant strategies, L3 resists invasion attempts if and only if $2R > T + S$, which always holds across cooperative dilemmas (6).

All four conditions hold across prisoner's dilemmas, so the corresponding stability index converges to one. In contrast, $P > S$ fails for every snowdrift game and $T > R$ fails for every stag hunt; the stability indices for these two classes thus converge to zero. $\square$

3.7.4 L4 (DISC, SS-SJ)

L4 belongs to $\mathcal{R}' \setminus \mathcal{R}''$ (Proposition 19); its assessment rule has $d_{\text{GBC}*} = \text{G}$ and $d_{\text{BBC}*} = \text{B}$. The former exposes it, like L1 and L3, to invasion by *CCC mutants (among them ALLC) under stag-hunt payoffs; the latter softens the vulnerability to CDISC present in L3, restricting it to the snowdrift games for which $R + P < 2S$ holds. The following proposition quantifies the resulting stability indices.

Proposition 32 (Stability of L4). *For $\epsilon = 0$ and in the limit $\chi \rightarrow 0$, L4 is stable across prisoner's dilemmas, but it can be invaded by CDISC if the cooperative dilemma is a snowdrift game for which $R + P < 2S$ holds, and it can always be invaded by $\tilde{p} = \text{*CCC}$ if the cooperative dilemma is a stag hunt. Thus:*

$$\lim_{\chi \rightarrow 0} \left(\sigma_{\text{L4}}^{\text{PD}} \right)_{\epsilon=0} = 1, \quad \lim_{\chi \rightarrow 0} \left(\sigma_{\text{L4}}^{\text{SG}} \right)_{\epsilon=0} = 2/3, \quad \lim_{\chi \rightarrow 0} \left(\sigma_{\text{L4}}^{\text{SH}} \right)_{\epsilon=0} = 0.$$

Proof. We compute $\Delta w(\chi)$ for each of the 15 mutant strategies $\tilde{p} \neq p$; the results are collected in Table 18.

A direct inspection of Table 18 reveals four conditions for L4 to resist invasion by a mutant strategy $\tilde{p} \neq p$:

1. If $\tilde{p} = *D**$, then L4 resists invasion attempts if and only if $R > P$, which always holds across cooperative dilemmas (3).
2. If $\tilde{p} = *CCC$, then L4 resists invasion attempts if and only if $T > R$, which holds for all prisoner's dilemmas (9) and snowdrift games (10), but for no stag hunt (11).
3. If $\tilde{p} = CCDC$ (i.e., CDISC), then L4 resists invasion attempts if and only if $R + P > 2S$, which holds for all prisoner's dilemmas (9) and stag hunts (11), but not for all snowdrift games (10).
4. For all remaining mutant strategies, L4 resists invasion attempts if and only if $2R > T + S$, which always holds across cooperative dilemmas (6).

All four conditions hold across prisoner's dilemmas, so the corresponding stability index converges to one. Within snowdrift games, the only binding condition is $R + P > 2S$: it fails precisely for the games with $R + P < 2S$, in which CDISC invades, and by Lemma 1 the snowdrift games satisfying $R + P > 2S$ make up a fraction $2/3$ of the class. Finally, $T > R$ fails for every stag hunt, in which the mutants of item 2 invade; the stability index for this class thus converges to zero. $\square$

3.7.5 L5 (DISC, SJ-SS)

L5 belongs to $\mathcal{R}'' \setminus \mathcal{R}'''$ (Proposition 19); its assessment rule has $d_{GBC*} = B$ and $d_{BBC*} = G$. The former removes the vulnerability to $*CCC$ mutants present in L1, L3, and L4 (the binding condition becomes $T > S$ rather than $T > R$); the latter exposes it, like L3, to invasion by CDISC under snowdrift payoffs. The following proposition quantifies the resulting stability indices.

Proposition 33 (Stability of L5). *For $\epsilon = 0$ and in the limit $\chi \rightarrow 0$, L5 is stable across prisoner's dilemmas and stag hunts, but it can always be invaded by CDISC if the cooperative dilemma is a snowdrift game. Thus:*

$$\lim_{\chi \rightarrow 0} \left(\sigma_{L5}^{PD} \right)_{\epsilon=0} = 1, \quad \lim_{\chi \rightarrow 0} \left(\sigma_{L5}^{SG} \right)_{\epsilon=0} = 0, \quad \lim_{\chi \rightarrow 0} \left(\sigma_{L5}^{SH} \right)_{\epsilon=0} = 1.$$

Proof. We compute $\Delta w(\chi)$ for each of the 15 mutant strategies $\tilde{p} \neq p$; the results are collected in Table 19.

A direct inspection of Table 19 reveals four conditions for L5 to resist invasion by a mutant strategy $\tilde{p} \neq p$:

1. If $\tilde{p} = *D**$, then L5 resists invasion attempts if and only if $R > P$, which always holds across cooperative dilemmas (3).
2. If $\tilde{p} = *CCC$, then L5 resists invasion attempts if and only if $T > S$, which always holds across cooperative dilemmas (12).
3. If $\tilde{p} = CCDC$ (i.e., CDISC), then L5 resists invasion attempts if and only if $P > S$, which holds for all prisoner's dilemmas (9) and stag hunts (11), but for no snowdrift game (10).
4. For all remaining mutant strategies, L5 resists invasion attempts if and only if $2R > T + S$, which always holds across cooperative dilemmas (6).

All four conditions hold across prisoner's dilemmas and stag hunts, so the corresponding stability indices converge to one. In contrast, $P > S$ fails for every snowdrift game, in which CDISC invades; the stability index for this class thus converges to zero. $\square$

Table 19: Taylor series expansion of the difference between resident and mutant expected payoffs for the leading-eight moral system L5 (DISC, SJ-SS) for different mutant strategies $\tilde{p} \neq \text{DISC}$. Here, no execution errors are considered ($\epsilon = 0$).

$\tilde{p}$ (ID)	$\tilde{p}$ (bitstring)	$\tilde{p}$ (name)	$\Delta w(\chi)$
0	DDDD	ALLD	$R - P + O(\chi)$
1	CDDD		$R - P + O(\chi)$
2	DCDD		$\frac{1}{2}(2R - T - S) + O(\chi)$
3	CCDD		$\frac{1}{2}(2R - T - S) + O(\chi)$
4	DDCD		$R - P + O(\chi)$
5	CDCD		$R - P + O(\chi)$
6	DCCD		$\frac{1}{2}(2R - T - S) + O(\chi)$
7	CCCD		$\frac{1}{2}(2R - T - S) + O(\chi)$
8	DDDC	PDISC	$\frac{1}{3}(R - P) + O(\chi)$
9	CDDC	ODISC	$\frac{1}{3}(R - P) + O(\chi)$
10	DCDC	DISC	—
11	CCDC	CDISC	$(P - S)\chi^2 + O(\chi^3)$
12	DDCC	ALLC	$\frac{1}{2}(R - P) + O(\chi)$
13	CDCC		$\frac{1}{2}(R - P) + O(\chi)$
14	DCCC		$(T - S)\chi + O(\chi^2)$
15	CCCC		$(T - S)\chi + O(\chi^2)$

3.7.6 L6 (DISC, Stern Judging)

L6 belongs to $\mathcal{Q}_{\text{DISC}}$ (cf. Eqs. (56) and (44)), the set of consistent discriminating systems with DISC as action rule. Its Stern Judging assessment rule has $\mathbf{d}_{\text{GBC}*} = \mathbf{B}$ and $\mathbf{d}_{\text{BBC}*} = \mathbf{B}$: like L5, it is protected against *CCC mutants; like L4, it can be invaded by CDISC only in the snowdrift games for which $R + P < 2S$ holds.

Proposition 34 (Stability of L6). *For $\epsilon = 0$ and in the limit $\chi \rightarrow 0$, L6 is stable across prisoner's dilemmas and stag hunts, but it can be invaded by CDISC if the cooperative dilemma is a snowdrift game for which $R + P < 2S$ holds. Thus:*

$$\lim_{\chi \rightarrow 0} \left(\sigma_{\text{L6}}^{\text{PD}} \right)_{\epsilon=0} = 1, \quad \lim_{\chi \rightarrow 0} \left(\sigma_{\text{L6}}^{\text{SG}} \right)_{\epsilon=0} = 2/3, \quad \lim_{\chi \rightarrow 0} \left(\sigma_{\text{L6}}^{\text{SH}} \right)_{\epsilon=0} = 1.$$

Proof. Immediate from Proposition 26, since L6 belongs to $\mathcal{Q}_{\text{DISC}}$. □

3.7.7 L7 (DISC, Staying)

L7 belongs to $\mathcal{R}' \setminus \mathcal{R}''$ (Proposition 19); its Staying assessment rule has $\mathbf{d}_{\text{GBC}*} = \mathbf{G}$ and $\mathbf{d}_{\text{BBD}*} = \mathbf{d}_{\text{BBC}*} = \mathbf{B}$. The former exposes it, like L1, L3, and L4, to invasion by *CCC mutants (among them ALLC) under stag-hunt payoffs. Because of the latter, a CDISC mutant — which deviates from DISC only by cooperating in interactions between bad individuals — incurs no reputational penalty for the deviation, so that L7 shares L3's vulnerability to CDISC under snowdrift payoffs despite having $\mathbf{d}_{\text{BBC}*} = \mathbf{B}$ like L4 and L6. The following proposition quantifies the resulting stability indices.

Table 20: Taylor series expansion of the difference between resident and mutant expected payoffs for the leading-eight moral system L7 (DISC, Staying) for different mutant strategies $\tilde{p} \neq \text{DISC}$. Here, no execution errors are considered ($\epsilon = 0$).

$\tilde{p}$ (ID)	$\tilde{p}$ (bitstring)	$\tilde{p}$ (name)	$\Delta w(\chi)$
0	DDDD	ALLD	$R - P + O(\chi)$
1	CDDD		$R - P + O(\chi)$
2	DCDD		$\frac{1}{2}(2R - T - S) + O(\chi)$
3	CCDD		$\frac{1}{2}(2R - T - S) + O(\chi)$
4	DDCD		$R - P + O(\chi)$
5	CDCD		$R - P + O(\chi)$
6	DCCD		$\frac{1}{2}(2R - T - S) + O(\chi)$
7	CCCD		$\frac{1}{2}(2R - T - S) + O(\chi)$
8	DDDC	PDISC	$\frac{1}{2}(R - P) + O(\chi)$
9	CDDC	ODISC	$\frac{1}{2}(R - P) + O(\chi)$
10	DCDC	DISC	—
11	CCDC	CDISC	$(P - S)\chi^2 + O(\chi^3)$
12	DDCC		$\frac{1}{2}(R - P) + O(\chi)$
13	CDCC		$\frac{1}{2}(R - P) + O(\chi)$
14	DCCC		$(T - R)\chi + O(\chi^2)$
15	CCCC	ALLC	$(T - R)\chi + O(\chi^2)$

Proposition 35 (Stability of L7). *For $\epsilon = 0$ and in the limit $\chi \rightarrow 0$, L7 is stable across prisoner's dilemmas, but it can always be invaded by CDISC if the cooperative dilemma is a snowdrift game, and by $\tilde{p} = \text{*CCC}$ if the cooperative dilemma is a stag hunt. Thus:*

$$\lim_{\chi \rightarrow 0} \left(\sigma_{L7}^{\text{PD}} \right)_{\epsilon=0} = 1, \quad \lim_{\chi \rightarrow 0} \left(\sigma_{L7}^{\text{SG}} \right)_{\epsilon=0} = 0, \quad \lim_{\chi \rightarrow 0} \left(\sigma_{L7}^{\text{SH}} \right)_{\epsilon=0} = 0.$$

Proof. We compute $\Delta w(\chi)$ for each of the 15 mutant strategies $\tilde{p} \neq p$; the results are collected in Table 20.

A direct inspection of Table 20 reveals four conditions for L7 to resist invasion by a mutant strategy $\tilde{p} \neq p$:

1. If $\tilde{p} = \text{*D**}$, then L7 resists invasion attempts if and only if $R > P$, which always holds across cooperative dilemmas (3).
2. If $\tilde{p} = \text{*CCC}$, then L7 resists invasion attempts if and only if $T > R$, which holds for all prisoner's dilemmas (9) and snowdrift games (10), but for no stag hunt (11).
3. If $\tilde{p} = \text{CCDC}$ (i.e., CDISC), then L7 resists invasion attempts if and only if $P > S$, which holds for all prisoner's dilemmas (9) and stag hunts (11), but for no snowdrift game (10).
4. For all remaining mutant strategies, L7 resists invasion attempts if and only if $2R > T + S$, which always holds across cooperative dilemmas (6).

All four conditions hold across prisoner's dilemmas, so the corresponding stability index converges to one. In contrast, $P > S$ fails for every snowdrift game and $T > R$ fails for every stag hunt; the stability indices for these two classes thus converge to zero. $\square$

Table 21: Taylor series expansion of the difference between resident and mutant expected payoffs for the leading-eight moral system L8 (DISC, Judging) for different mutant strategies $\tilde{p} \neq \text{DISC}$. Here, no execution errors are considered ($\epsilon = 0$).

$\tilde{p}$ (ID)	$\tilde{p}$ (bitstring)	$\tilde{p}$ (name)	$\Delta w(\chi)$
0	DDDD	ALLD	$R - P + O(\chi)$
1	CDDD		$R - P + O(\chi)$
2	DCDD		$\frac{1}{2}(2R - T - S) + O(\chi)$
3	CCDD		$\frac{1}{2}(2R - T - S) + O(\chi)$
4	DDCD		$R - P + O(\chi)$
5	CDCD		$R - P + O(\chi)$
6	DCCD		$\frac{1}{2}(2R - T - S) + O(\chi)$
7	CCCD		$\frac{1}{2}(2R - T - S) + O(\chi)$
8	DDDC	PDISC	$\frac{1}{2}(R - P) + O(\chi)$
9	CDDC	ODISC	$\frac{1}{2}(R - P) + O(\chi)$
10	DCDC	DISC	—
11	CCDC	CDISC	$(P - S)\chi^2 + O(\chi^3)$
12	DDCC		$\frac{2}{3}(R - P) + O(\chi)$
13	CDCC		$\frac{2}{3}(R - P) + O(\chi)$
14	DCCC		$(T - S)\chi + O(\chi^2)$
15	CCCC	ALLC	$(T - S)\chi + O(\chi^2)$

3.7.8 L8 (DISC, Judging)

L8 belongs to $\mathcal{R}'' \setminus \mathcal{R}'''$ (Proposition 19); its Judging assessment rule has $\mathbf{d}_{\text{GBC}^*} = \mathbf{B}$ and $\mathbf{d}_{\text{BBD}^*} = \mathbf{d}_{\text{BBC}^*} = \mathbf{B}$. Like L5, it is protected against *CCC mutants by its first bit; like L7, its two remaining bits leave a CDISC mutant's deviation without reputational penalty, so that it can be invaded by CDISC under snowdrift payoffs. The following proposition quantifies the resulting stability indices.

Proposition 36 (Stability of L8). *For $\epsilon = 0$ and in the limit $\chi \rightarrow 0$, L8 is stable across prisoner's dilemmas and stag hunts, but it can always be invaded by CDISC if the cooperative dilemma is a snowdrift game. Thus:*

$$\lim_{\chi \rightarrow 0} \left(\sigma_{\text{L8}}^{\text{PD}} \right)_{\epsilon=0} = 1, \quad \lim_{\chi \rightarrow 0} \left(\sigma_{\text{L8}}^{\text{SG}} \right)_{\epsilon=0} = 0, \quad \lim_{\chi \rightarrow 0} \left(\sigma_{\text{L8}}^{\text{SH}} \right)_{\epsilon=0} = 1.$$

Proof. We compute $\Delta w(\chi)$ for each of the 15 mutant strategies $\tilde{p} \neq p$; the results are collected in Table 21.

A direct inspection of Table 21 reveals four conditions for L8 to resist invasion by a mutant strategy $\tilde{p} \neq p$:

1. If $\tilde{p} = \text{*D**}$, then L8 resists invasion attempts if and only if $R > P$, which always holds across cooperative dilemmas (3).
2. If $\tilde{p} = \text{*CCC}$, then L8 resists invasion attempts if and only if $T > S$, which always holds across cooperative dilemmas (12).
3. If $\tilde{p} = \text{CCDC}$ (i.e., CDISC), then L8 resists invasion attempts if and only if $P > S$, which holds for all prisoner's dilemmas (9) and stag hunts (11), but for no snowdrift game (10).

Table 22: Stability of the leading eight across cooperative dilemmas, for $\epsilon = 0$ and in the limit $\chi \rightarrow 0$. Each entry gives the limiting value of the stability index for the indicated game class: a value of 1 indicates stability across the entire class, a value of $2/3$ stability across a strict subset, and a value of 0 instability throughout. The last column gives the average stability index σ (32).

Moral system	Action rule	Assessment rule	σ^{PD}	σ^{SG}	σ^{SH}	σ
L1	CDISC	Standing	1	1	0	$2/3$
L2	CDISC	Consistent Standing	1	1	1	1
L3	DISC	Simple Standing	1	0	0	$1/3$
L4	DISC	SS-SJ	1	$2/3$	0	$5/9$
L5	DISC	SJ-SS	1	0	1	$2/3$
L6	DISC	Stern Judging	1	$2/3$	1	$8/9$
L7	DISC	Staying	1	0	0	$1/3$
L8	DISC	Judging	1	0	1	$2/3$

4. For all remaining mutant strategies, L8 resists invasion attempts if and only if $2R > T + S$, which always holds across cooperative dilemmas (6).

All four conditions hold across prisoner’s dilemmas and stag hunts, so the corresponding stability indices converge to one. In contrast, $P > S$ fails for every snowdrift game, in which CDISC invades; the stability index for this class thus converges to zero. $\square$

3.7.9 Summary: the leading eight

Table 22 collects the stability results for the eight leading-eight moral systems (Propositions 29–36). L2 (CDISC, Consistent Standing) is the only universally stable system; every other member can be invaded in at least one class of cooperative dilemma. All six systems with DISC as action rule can be invaded by CDISC in snowdrift games — in every such game for L3, L5, L7, and L8, and in those satisfying $R + P < 2S$ for L4 and L6. The reason is that $S > P$ holds in snowdrift games: when meeting a co-player in bad standing while in bad standing itself, a contrite discriminator cooperates (earning S) where the resident defects (earning P), and thus profits from the deviation. The four systems with $d_{\text{GBC}^*} = G$ (L1, L3, L4, and L7) — among them L1, which pairs the CDISC action rule with the inconsistent Standing norm — fail in every stag hunt, where $R > T$ holds: mutants that also cooperate with co-players in bad standing retain their good standing and earn R where residents earn only T . These results reinforce the conclusion of the preceding sections: among the leading eight, universal stability requires both consistency and contrition, a combination found only in L2.

4 Numerical results

The analytical results of the preceding sections were derived in the idealized regime of zero execution errors ($\epsilon = 0$) and vanishing assignment errors ($\chi \rightarrow 0$), and the numerical results of the main text were computed for a single combination of small error rates ($\epsilon = \chi = 0.02$). This section presents numerical results verifying that our findings are robust to these choices. We first count moral systems up to the mirror symmetry of Section 2.5 and specify the representatives our numerical explorations evaluate (Section 4.1), and then detail the computational pipeline behind

these numerical results (Section 4.2). We finally examine how the stability and cooperation indices behave across a grid of error-rate combinations (Section 4.3), how stability is distributed across the three classes of cooperative dilemma (Section 4.4), and how results change under the alternative, area-weighted construction of the average stability index (Section 4.5).

4.1 Mirror classes and representatives

Section 2.5 established that a moral system and its mirror are behaviorally equivalent: they share the same cooperation and stability indices (Proposition 2), whereas label-dependent properties such as self-approval hold for at most one member of each mirror pair (Corollary 1). Here we extend these results to consistency, count moral systems up to the mirror symmetry, and specify the representatives evaluated in our numerical explorations.

Consistency is a label-dependent property: it requires conformity with the resident prescription to be rewarded with the particular label G. Unsurprisingly, it is not mirror-invariant.

Proposition 37 (Mirrors of consistent systems). *Let (p, d) be a consistent system. Then its mirror $(\hat{p}, \hat{d})$ satisfies*

$$\hat{d}(r_1, r_2, \hat{p}(r_1, r_2), \hat{p}(r_2, r_1)) = B, \text{ and} \quad (128a)$$

$$\hat{d}(r_1, r_2, \neg \hat{p}(r_1, r_2), \hat{p}(r_2, r_1)) = G \quad (128b)$$

for all $r_1, r_2 \in \{B, G\}$. Consequently, the mirror of a consistent system is not consistent, no consistent system is a self-mirror, and each mirror class contains at most one consistent system.

Proof. By (61), $\hat{d}(r_1, r_2, \hat{p}(r_1, r_2), \hat{p}(r_2, r_1)) = \neg d(\neg r_1, \neg r_2, p(\neg r_1, \neg r_2), p(\neg r_2, \neg r_1)) = \neg G = B$, using the consistency condition (43a) at the reputation profile $(\neg r_1, \neg r_2)$; this establishes Eq. (128a), and Eq. (128b) follows in the same way from (43b). Since consistency (43a) requires the left-hand side of Eq. (128a) to equal G for all reputation profiles, $(\hat{p}, \hat{d})$ is not consistent. If a consistent system were a self-mirror, its mirror – the system itself – would be consistent, contradicting the non-consistency of the mirror just established; the same argument shows that the two members of a mirror pair cannot both be consistent. $\square$

In words, Eq. (128) says that the mirror of a consistent system is consistent with the labels swapped: it rewards conformity with the label B, which plays for the mirror system the role that G plays for the original one. This is precisely the sense in which the two systems are relabelings of one another.

We now count self-mirrors, and with them the mirror classes.

Proposition 38 (Number of self-mirrors). *A moral system (p, d) is a self-mirror if and only if both*

$$p_{BB} = p_{GG} \text{ and } p_{BG} = p_{GB}, \text{ and} \quad (129a)$$

$$d_{r_1 r_2 a_1 a_2} = \neg d_{\neg r_1 \neg r_2 a_1 a_2} \text{ for all } r_1, r_2, a_1, a_2 \quad (129b)$$

hold. Exactly four strategies satisfy condition (129a) (the palindromic bitstrings DDDD, DCCD, CDDC, and CCCC) and exactly $2^8 = 256$ assessment rules satisfy condition (129b), so there are $4 \times 256 = 1,024$ self-mirrors.

Proof. Conditions (129a) and (129b) restate $\hat{p} = p$ and $\hat{d} = d$, respectively, using (61). For strategies, $\hat{p} = p$ says that the bitstring $p_{BB}p_{BG}p_{GB}p_{GG}$ is a palindrome, which leaves the two bits p_{BB} and p_{BG} free: the four solutions are DDDD (ALLD), DCCD, CDDC (ODISC), and CCCC (ALLC). For assessment rules, the map $(r_1, r_2, a_1, a_2) \mapsto (\neg r_1, \neg r_2, a_1, a_2)$ pairs the sixteen argument profiles into eight pairs (it has no fixed point, since $(r_1, r_2) \neq (\neg r_1, \neg r_2)$ always holds), and on each pair the constraint in (129b) admits exactly two of the four joint assignments; hence $2^8 = 256$ solutions. $\square$

There are exactly as many self-mirrors as consistent discriminating systems – 1,024 in both cases – but the coincidence is accidental: the former number arises as (four palindromic strategies) $\times$ (256 self-mirror assessment rules), the latter as (four discriminating strategies) $\times$ (256 consistent assessment rules per strategy), and the two sets are disjoint by Proposition 37. The two sets even share a strategy: ODISC ($p = \text{CDDC}$) is the unique discriminating strategy equal to its own mirror – yet, again by Proposition 37, no consistent ODISC system is a self-mirror.

Proposition 39 (Number of mirror classes). *Let $\mathcal{M}$ be a set of moral systems that is closed under the mirror map, and let m be the number of self-mirrors in $\mathcal{M}$. Then $\mathcal{M}$ partitions into*

$$\frac{|\mathcal{M}| - m}{2} + m = \frac{|\mathcal{M}| + m}{2} \quad (130)$$

mirror classes. In particular, the full space of $2^{20} = 1,048,576$ moral systems, which contains $m = 1,024$ self-mirrors (Proposition 38), partitions into $(2^{20} + 1,024) / 2 = 524,800$ mirror classes.

Proof. The mirror map restricts to an involution on $\mathcal{M}$, so it partitions $\mathcal{M}$ into classes of size one or two, and the classes of size one are exactly the self-mirrors in $\mathcal{M}$. The number of classes is therefore m plus half the number $|\mathcal{M}| - m$ of remaining systems. $\square$

Equation (130) is an instance of Burnside’s lemma (ref. 5, Chapter 10), applied to the two-element group generated by the mirror map: the number of orbits (the mirror classes) equals the average number of elements of $\mathcal{M}$ fixed by the group elements – $|\mathcal{M}|$ for the identity and m for the mirror.

Beyond the full space, Eq. (130) applies to any mirror-closed subset, such as those obtained by pairing the 2^4 strategies with a type class $\mathcal{T}_{\mathcal{U}}$ of assessment rules (Section 1.5): the mirror of a rule of type $\mathcal{U}$ is again of type $\mathcal{U}$, since the mirror negates reputation arguments and outputs but never adds a dependence on a new argument. If $\mathcal{U}$ contains at least one reputation argument, the negation $(r_1, r_2) \mapsto (\neg r_1, \neg r_2)$ pairs the $2^{|\mathcal{U}|}$ argument profiles of type $\mathcal{U}$ into $2^{|\mathcal{U}|-1}$ pairs, so that, by the argument in the proof of Proposition 38, exactly $2^{2^{|\mathcal{U}|-1}}$ rules of type $\mathcal{U}$ are their own mirrors; if $\mathcal{U}$ contains no reputation argument, a self-mirror rule would have to equal its own negation, so there is none. For instance, the space explored by Ohtsuki and Iwasa² pairs the 2^4 strategies with the $2^{2^3} = 256$ third-order rules of type $a_1 r_1 r_2$: it contains $4 \times 2^{2^2} = 2^6$ self-mirrors (four palindromic strategies times $2^{2^2} = 16$ self-mirror rules), so Eq. (130) yields $(2^{12} + 2^6) / 2 = 2,080$ mirror classes, recovering the count of ‘effective pairs’ derived in their Appendix A.

Because both members of a mirror pair share the same cooperation and stability indices (Proposition 2), numerical explorations of the space of moral systems need to evaluate only one member per mirror class – a reduction already exploited by Ohtsuki and Iwasa², who report a single member of each pair and omit its mirror image. Sweeping the full space of moral systems, however, requires an explicit selection rule, which the following definition provides.

Definition 16 (Majority-good representatives). *Fix error rates $\chi \in (0, 1/2)$ and $\epsilon \in [0, 1/2)$. The representative of a mirror class is: its unique member, if the class is a self-mirror; otherwise, the member with $g_{p,d}^* > 1/2$, if there is one; otherwise (that is, if both members of the pair satisfy $g_{p,d}^* = 1/2$), the member with the larger strategy ID (Table 1) or, if the two members share the same strategy, the member whose assessment rule has the larger ID (the sixteen bits $\mathbf{d}_{r_1 r_2 a_1 a_2}$ read as a binary number, in the same little-endian convention as in Table 1).*

Definition 16 is exhaustive and unambiguous: by Lemma 3, the two members of a mirror pair have resident reputation equilibria g^* and $1 - g^*$, so either exactly one member satisfies $g^* > 1/2$ or both sit at $g^* = 1/2$, in which case their strategy IDs differ unless $\hat{p} = p$, and then their assessment-rule IDs differ. We call the representatives ‘majority-good’ because, in the generic case, the representative is the member of the pair whose residents mostly hold good reputations at the working error rates.

All numerical explorations reported in the main text and in this section evaluate one representative per mirror class; by Proposition 2, which member is evaluated is immaterial. Selecting majority-good representatives parallels, at finite error rates, the restriction to self-improving systems (Corollary 1): a self-improving system satisfies $g_{p,d}^* > 1/2$ for sufficiently small error rates, and is then the representative of its mirror class.

4.2 Computation of the cooperation and stability indices

We now describe the pipeline used to compute the cooperation and stability indices reported in the main text (Fig. 2; $\chi = \epsilon = 0.02$) and in the robustness analyses of this section. For each combination of error rates, we evaluate the 524,800 majority-good representatives of the mirror classes (Definition 16), obtained by enumerating all 2^{20} moral systems – with conditional or unconditional action rules – and retaining one representative per mirror class (Proposition 39).

For each representative (p, d) , we first compute the resident reputation equilibrium $g_{p,d}^*$, locating the unique root in $(0, 1)$ of the quadratic with coefficients (22) by bisection, and from it the cooperation index $\kappa_{p,d}$ (23).

The stability indices are estimated by Monte Carlo integration. For each game class $\Gamma \in \{\text{PD}, \text{SG}, \text{SH}\}$, we draw once 1,000 payoff configurations uniformly at random from the normalized payoff region $\mathcal{G}_\Gamma$ by rejection sampling: for prisoner’s dilemmas, $T = 1$ and $S = 0$ are fixed and (R, P) is drawn uniformly subject to (13); for snowdrift games, $T = 1$ and $P = 0$ are fixed and (R, S) is drawn uniformly subject to (14); and for stag hunts, $R = 1$ and $S = 0$ are fixed and (T, P) is drawn uniformly subject to (15). The resulting three samples are archived and reused for every moral system and every error-rate combination, so that comparisons across systems and error rates are based on the same payoff configurations. For each system and each payoff configuration, we then check the stability condition (31) directly: we compute the resident payoff $w_{p,d}$ (28) and the mutant payoffs $\tilde{w}_{\tilde{p},p,d}$ (30) for all 15 alternative strategies $\tilde{p} \neq p$, and record whether $w_{p,d} > \tilde{w}_{\tilde{p},p,d}$ holds for all of them. The estimate of $\sigma_{p,d}^\Gamma(\epsilon, \chi)$ is the fraction of the 1,000 payoff configurations at which (p, d) is stable – a binomial proportion, whose standard error $\sqrt{\sigma_{p,d}^\Gamma (1 - \sigma_{p,d}^\Gamma) / 1,000}$ is at most $\sqrt{1/(4 \times 1,000)} \approx 0.016$. The average stability index then follows from Eq. (32).

All computations were implemented in a dedicated Julia package accompanying this paper, with the archived payoff samples distributed alongside the code; evaluating all representatives takes on the order of one hour per game class and error-rate combination on a standard desktop

computer. The code, the archived payoff samples, and the computed indices are all openly available; see the Data availability and Code availability statements in the main text.

4.3 Alternative error rates

Fig. 5 shows the stability and cooperation indices of all moral systems for a grid of error-rate combinations, with execution error rates $\epsilon \in \{0, 0.01, 0.02, 0.1\}$ and assignment error rates $\chi \in \{0.01, 0.02, 0.05, 0.1\}$. Larger error rates lower both indices, but the cluster of consistent discriminating systems in the upper-right corner persists across all combinations examined, confirming that our results are not an artifact of the error rates chosen for the baseline analysis ($\epsilon = \chi = 0.02$).

Fig. 6 zooms in on the consistent discriminating systems, showing their stability and cooperation indices over the same grid of error rates so that the variation within the cluster becomes visible. L2 is among the most stable systems at every combination. At the largest assignment error rate we examine ($\chi = 0.1$), however, the most stable consistent PDISC systems match or slightly exceed it in stability – the differences lie at the resolution of our Monte Carlo estimates (Section 4.2) – while cooperating markedly less. Whatever the error rates, the consistent CDISC systems remain the most cooperative of the consistent discriminating systems.

For completeness, Figs. 7–22 reproduce Fig. 2 of the main text for each of the sixteen combinations of the error rates $\chi \in \{0.01, 0.02, 0.05, 0.1\}$ and $\epsilon \in \{0, 0.01, 0.02, 0.1\}$, with no jitter applied to the scatter points, so that small differences in stability and cooperation between individual systems can be read off reliably.

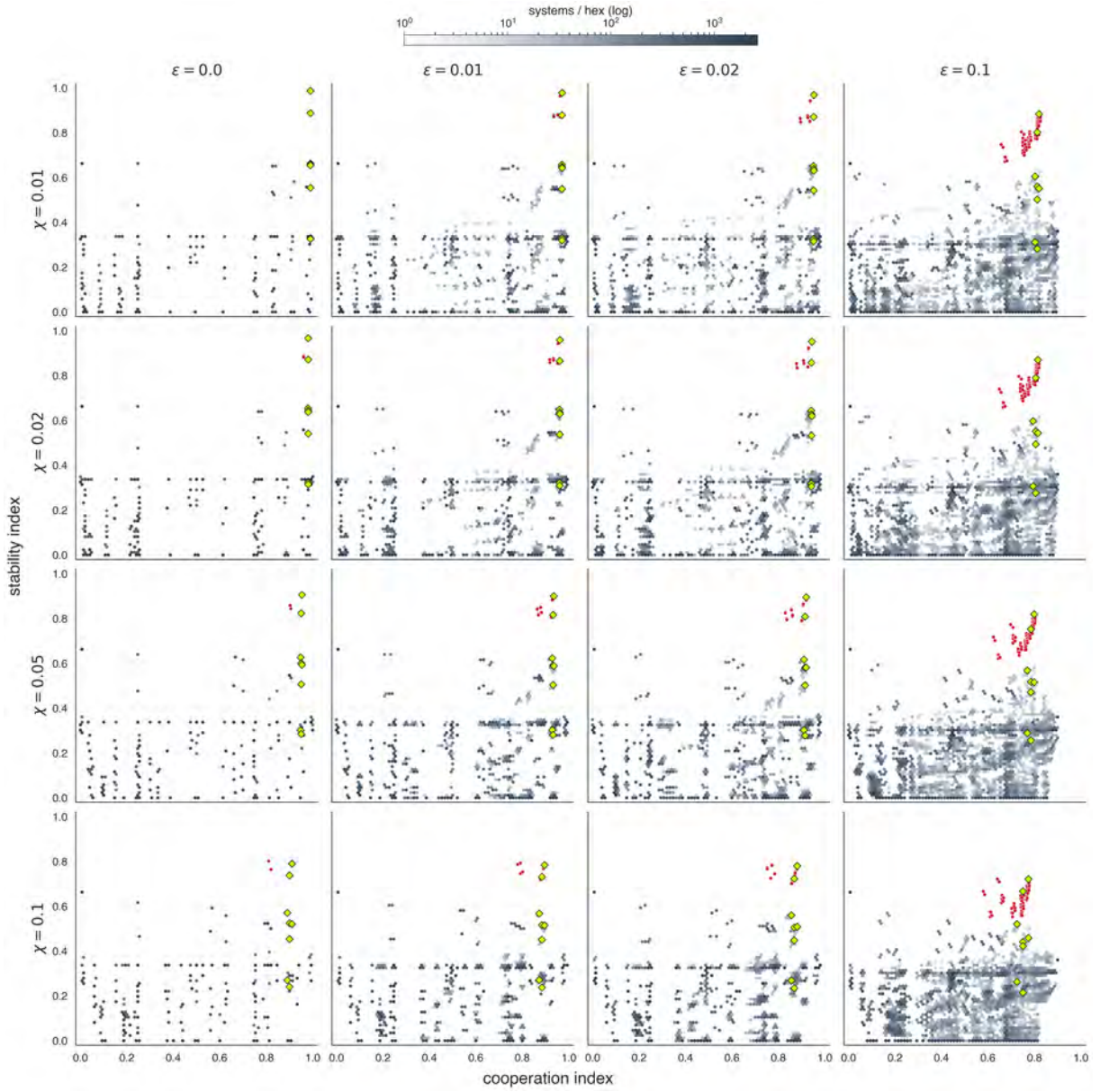


Figure 5: Stability and cooperation indices across a range of error rates. Each panel shows, for one combination of execution error rate $\epsilon \in \{0, 0.01, 0.02, 0.1\}$ (columns) and assignment error rate $\chi \in \{0.01, 0.02, 0.05, 0.1\}$ (rows), the cooperation index (κ) and average stability index (σ) of each of the 524,800 mirror-class representatives (Section 4.1). The consistent discriminating systems are shown as red circles, the leading eight as green diamonds, and the remaining systems as a hexbin density layer on a logarithmic color scale.

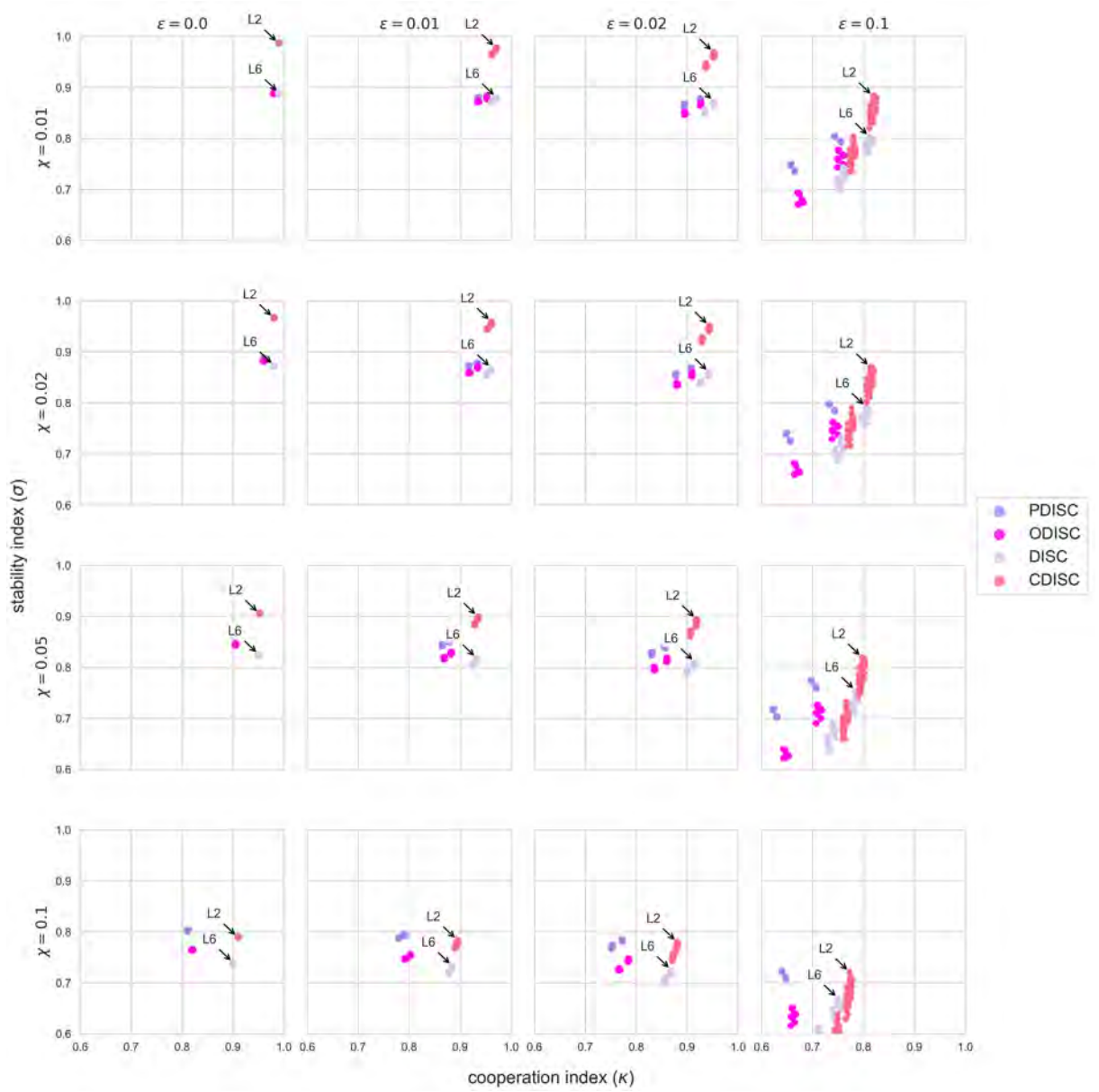


Figure 6: Zoomed scatter plots of the stability and cooperation indices of the consistent discriminating systems, for each combination of execution error rate $\epsilon \in \{0, 0.01, 0.02, 0.1\}$ (columns) and assignment error rate $\chi \in \{0.01, 0.02, 0.05, 0.1\}$ (rows), colored by strategy. L2 and L6, the only consistent discriminating systems among the leading eight, are marked with arrows. L2 is among the most stable systems at every combination. Points are jittered for clarity.

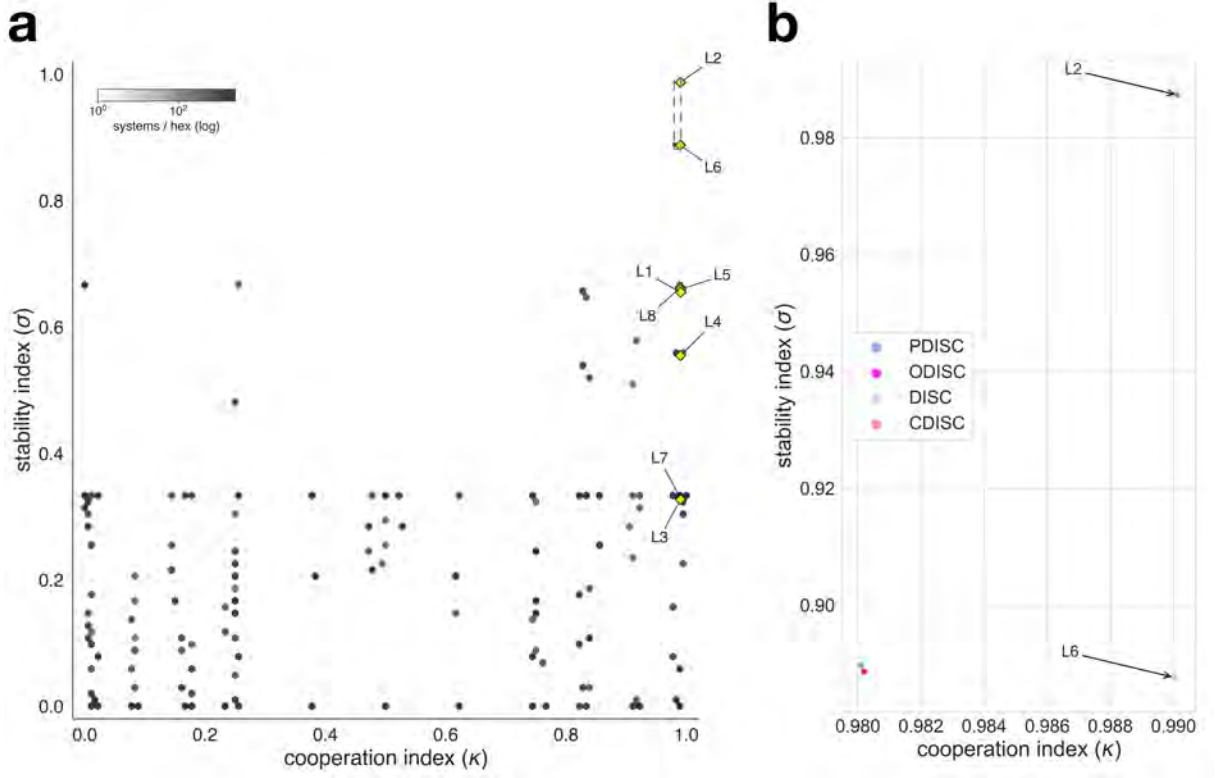


Figure 7: Stability and cooperation indices of all moral systems, without jitter, for $\chi = 0.01$ and $\epsilon = 0$. **a**, Each of the 524,800 moral systems that are distinct up to the relabeling of reputations (one representative per mirror class; Section 4.1) is positioned by its resident cooperation index (κ) and average stability index (σ); the consistent discriminating systems are shown as red circles, the leading eight as green diamonds, and the remaining systems as a hexbin density layer on a logarithmic color scale. **b**, Zoomed-in scatter plot of the consistent discriminating systems, colored by strategy (PDISC, ODISC, DISC, CDISC). No jitter is applied in either panel, so that small differences between individual systems can be read off reliably.

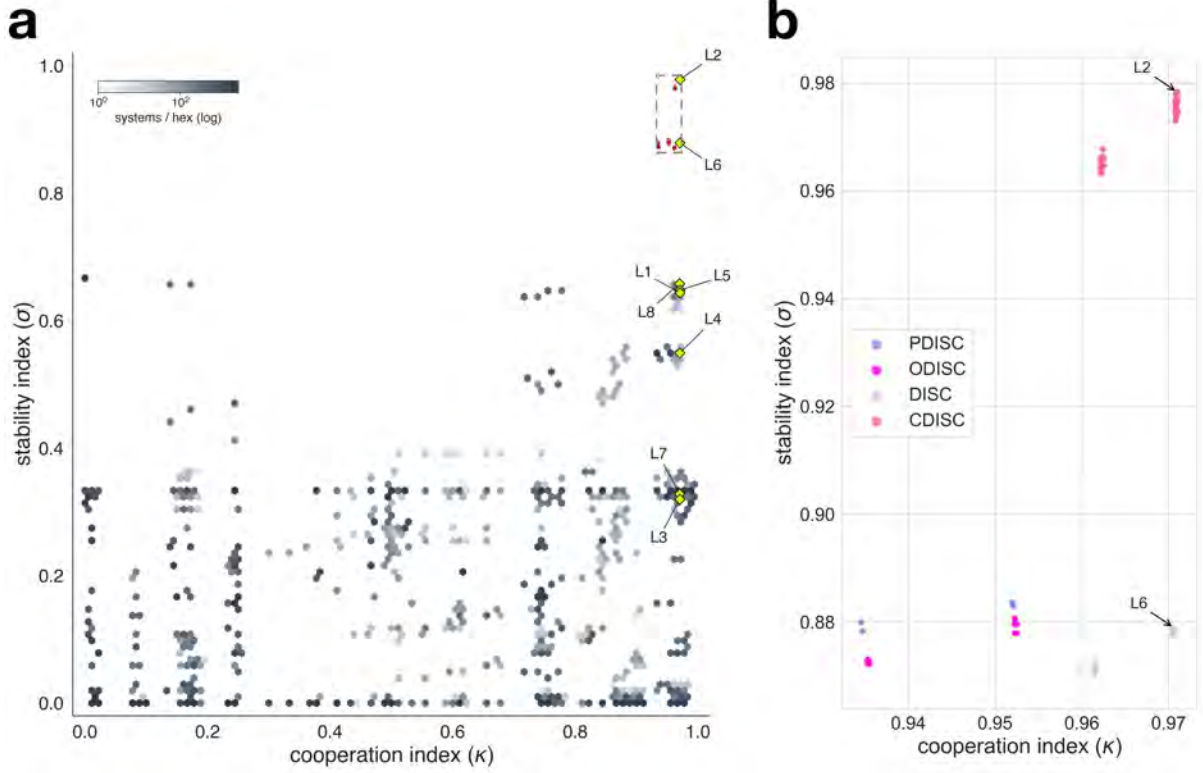


Figure 8: Stability and cooperation indices of all moral systems, without jitter, for $\chi = 0.01$ and $\epsilon = 0.01$. **a**, Each of the 524,800 moral systems that are distinct up to the relabeling of reputations (one representative per mirror class; Section 4.1) is positioned by its resident cooperation index (κ) and average stability index (σ); the consistent discriminating systems are shown as red circles, the leading eight as green diamonds, and the remaining systems as a hexbin density layer on a logarithmic color scale. **b**, Zoomed-in scatter plot of the consistent discriminating systems, colored by strategy (PDISC, ODISC, DISC, CDISC). No jitter is applied in either panel, so that small differences between individual systems can be read off reliably.

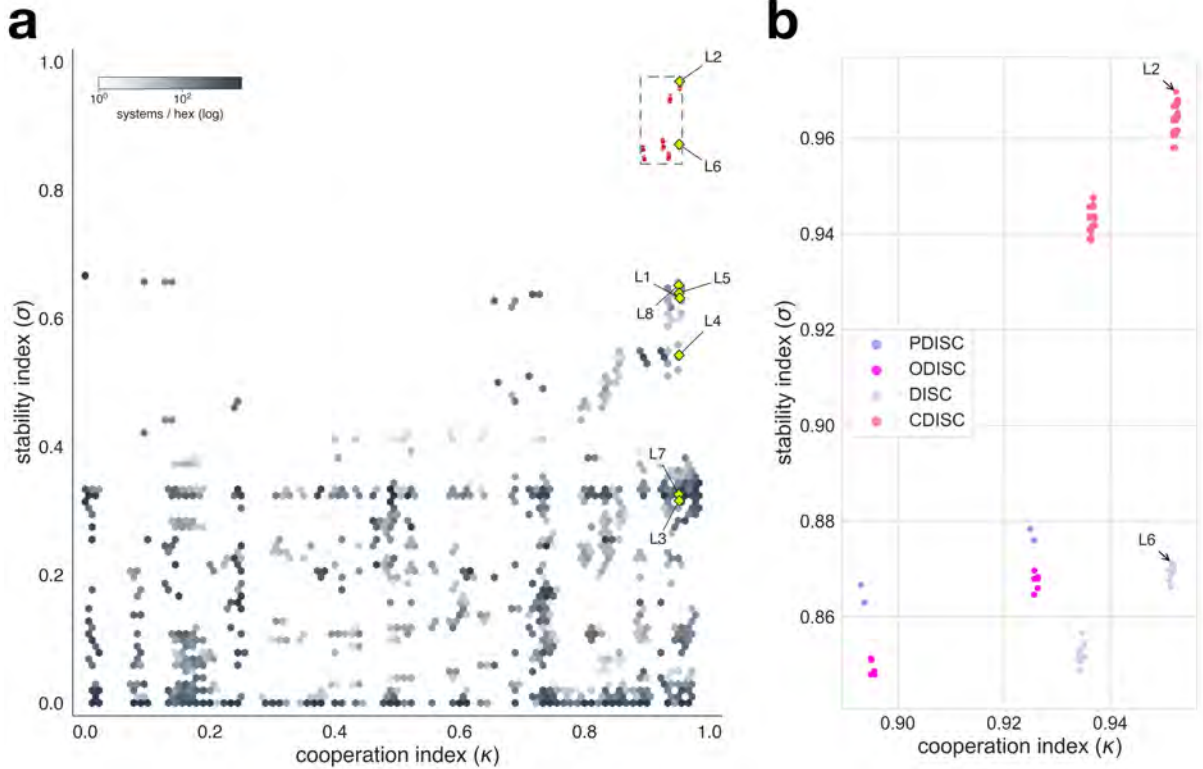


Figure 9: Stability and cooperation indices of all moral systems, without jitter, for $\chi = 0.01$ and $\epsilon = 0.02$. **a**, Each of the 524,800 moral systems that are distinct up to the relabeling of reputations (one representative per mirror class; Section 4.1) is positioned by its resident cooperation index (κ) and average stability index (σ); the consistent discriminating systems are shown as red circles, the leading eight as green diamonds, and the remaining systems as a hexbin density layer on a logarithmic color scale. **b**, Zoomed-in scatter plot of the consistent discriminating systems, colored by strategy (PDISC, ODISC, DISC, CDISC). No jitter is applied in either panel, so that small differences between individual systems can be read off reliably.

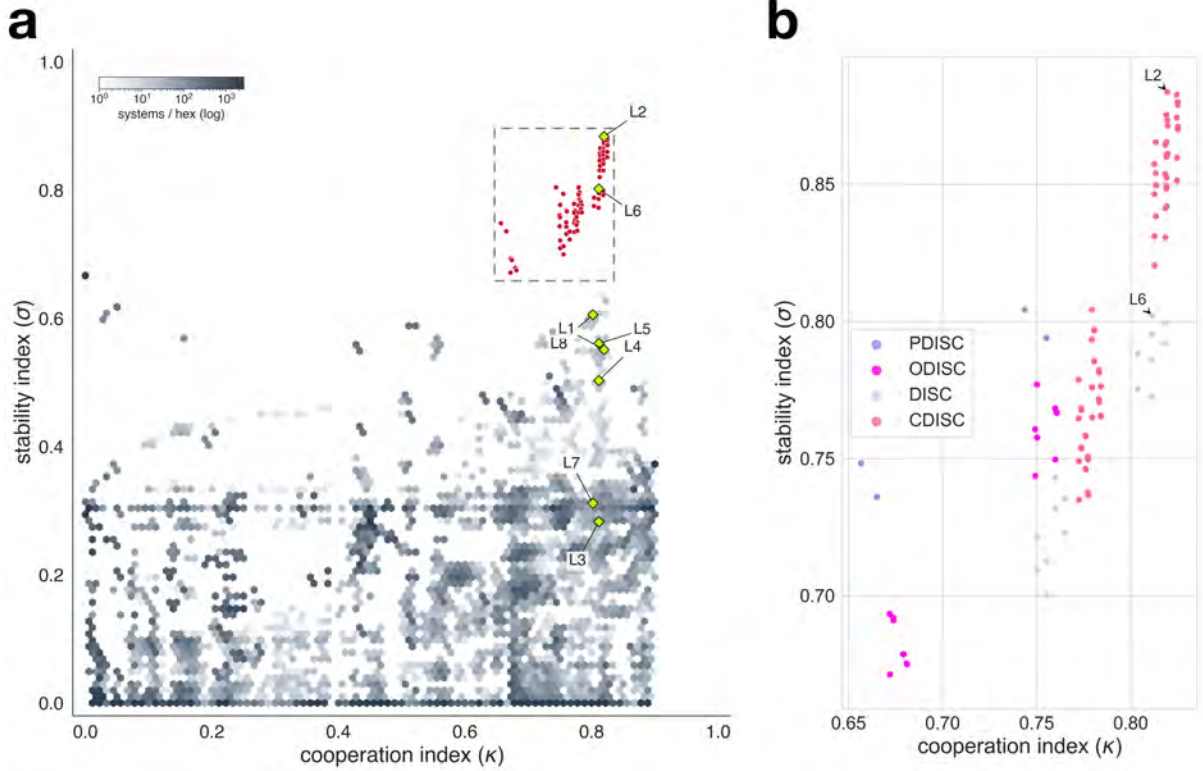


Figure 10: Stability and cooperation indices of all moral systems, without jitter, for $\chi = 0.01$ and $\epsilon = 0.1$. **a**, Each of the 524,800 moral systems that are distinct up to the relabeling of reputations (one representative per mirror class; Section 4.1) is positioned by its resident cooperation index (κ) and average stability index (σ); the consistent discriminating systems are shown as red circles, the leading eight as green diamonds, and the remaining systems as a hexbin density layer on a logarithmic color scale. **b**, Zoomed-in scatter plot of the consistent discriminating systems, colored by strategy (PDISC, ODISC, DISC, CDISC). No jitter is applied in either panel, so that small differences between individual systems can be read off reliably.

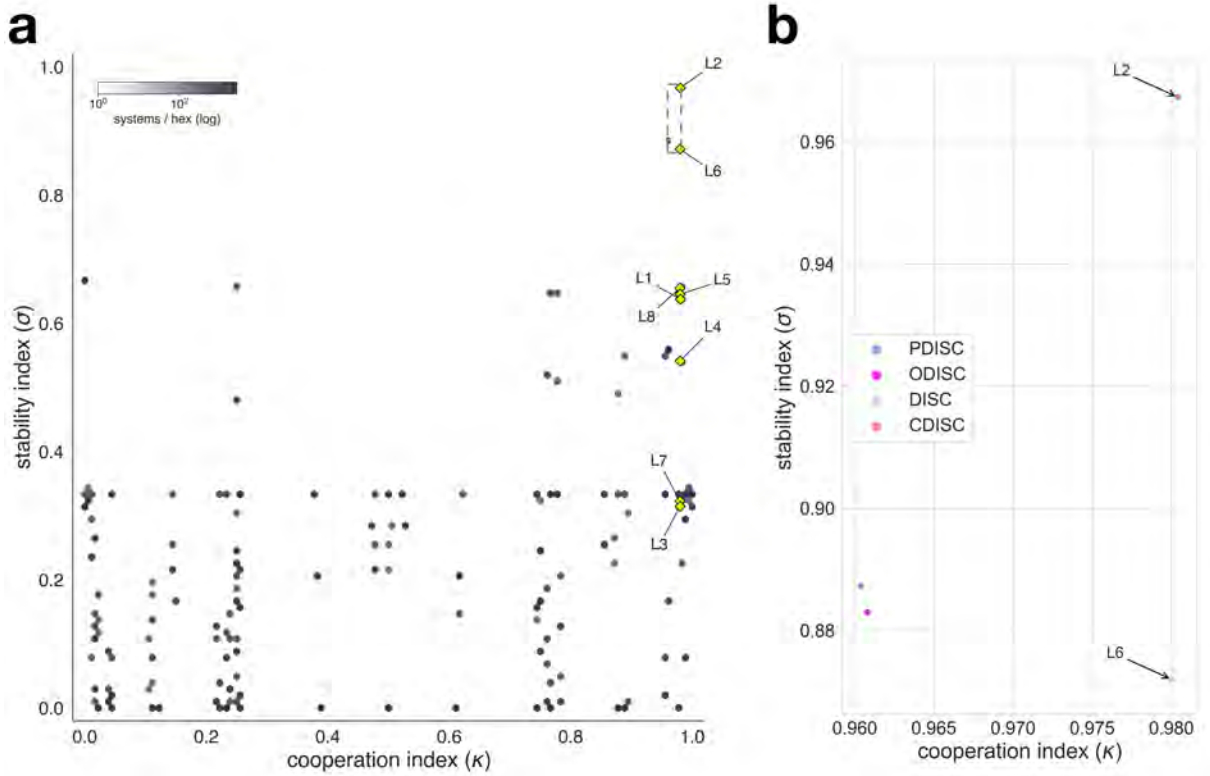


Figure 11: Stability and cooperation indices of all moral systems, without jitter, for $\chi = 0.02$ and $\epsilon = 0$. **a**, Each of the 524,800 moral systems that are distinct up to the relabeling of reputations (one representative per mirror class; Section 4.1) is positioned by its resident cooperation index (κ) and average stability index (σ); the consistent discriminating systems are shown as red circles, the leading eight as green diamonds, and the remaining systems as a hexbin density layer on a logarithmic color scale. **b**, Zoomed-in scatter plot of the consistent discriminating systems, colored by strategy (PDISC, ODISC, DISC, CDISC). No jitter is applied in either panel, so that small differences between individual systems can be read off reliably.

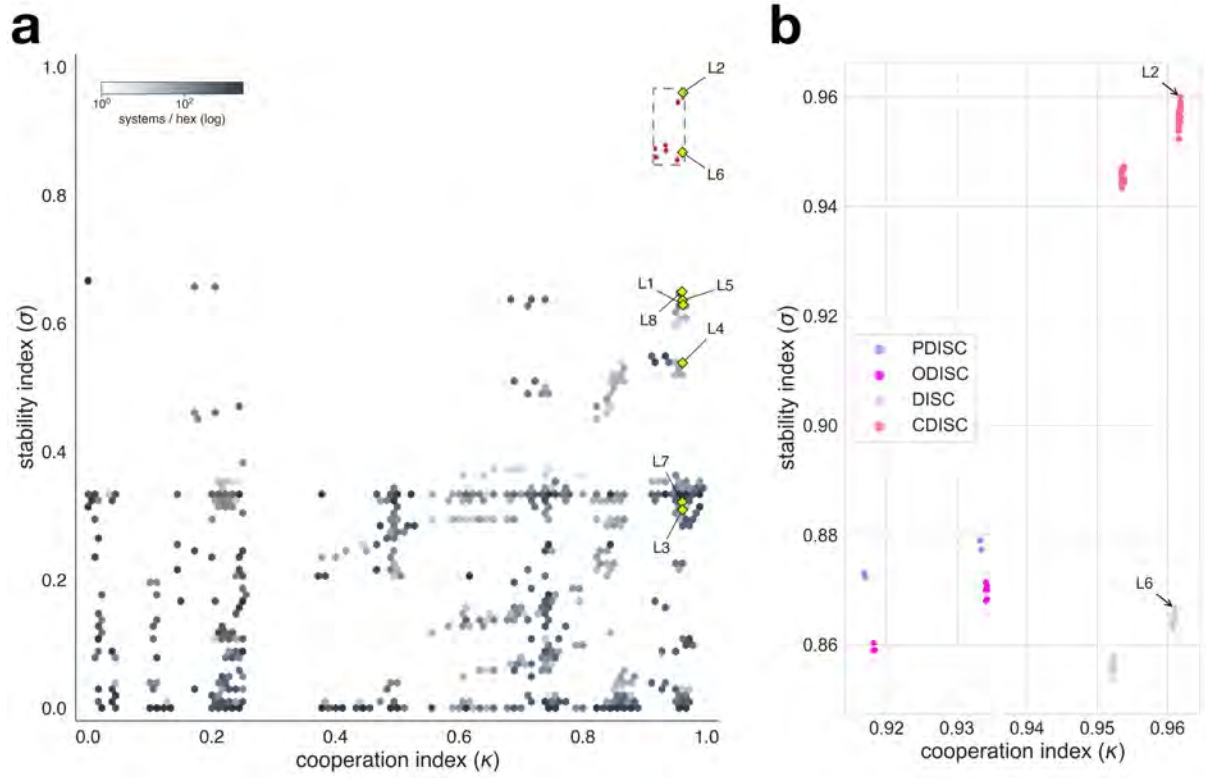


Figure 12: Stability and cooperation indices of all moral systems, without jitter, for $\chi = 0.02$ and $\epsilon = 0.01$. **a**, Each of the 524,800 moral systems that are distinct up to the relabeling of reputations (one representative per mirror class; Section 4.1) is positioned by its resident cooperation index (κ) and average stability index (σ); the consistent discriminating systems are shown as red circles, the leading eight as green diamonds, and the remaining systems as a hexbin density layer on a logarithmic color scale. **b**, Zoomed-in scatter plot of the consistent discriminating systems, colored by strategy (PDISC, ODISC, DISC, CDISC). No jitter is applied in either panel, so that small differences between individual systems can be read off reliably.

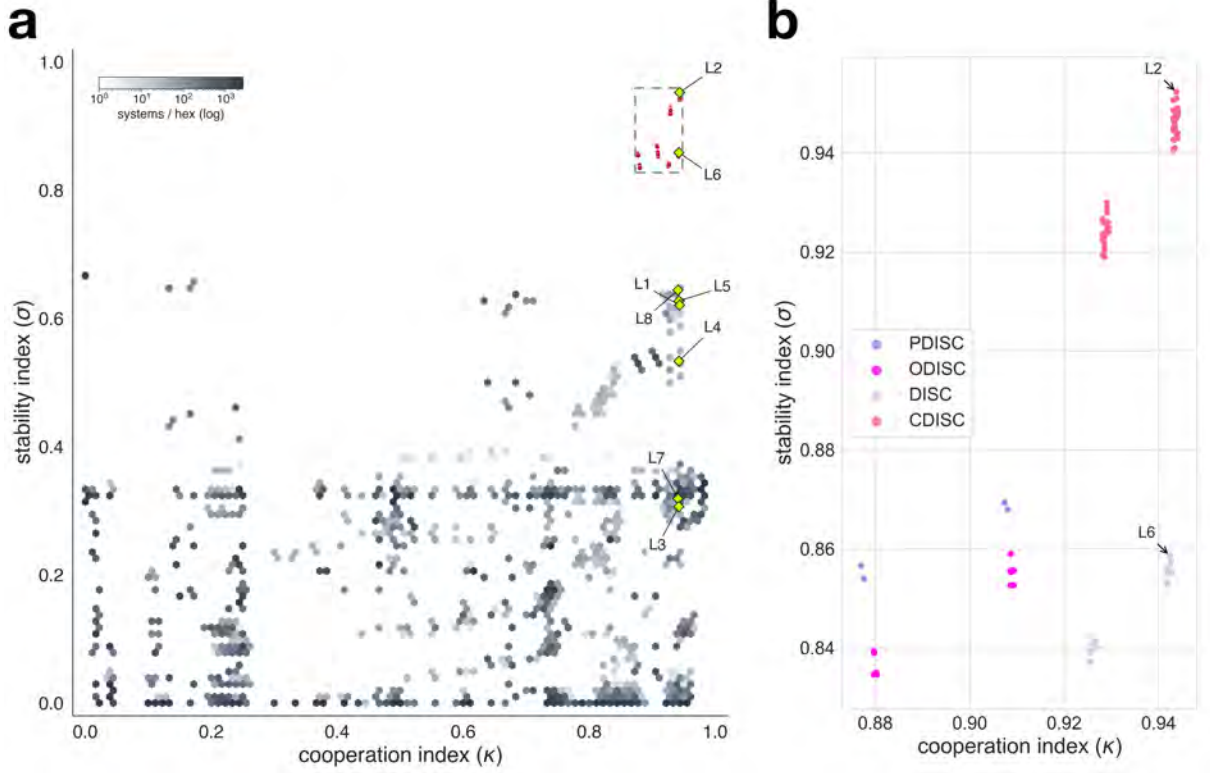


Figure 13: Stability and cooperation indices of all moral systems, without jitter, for $\chi = 0.02$ and $\epsilon = 0.02$. **a**, Each of the 524,800 moral systems that are distinct up to the relabeling of reputations (one representative per mirror class; Section 4.1) is positioned by its resident cooperation index (κ) and average stability index (σ); the consistent discriminating systems are shown as red circles, the leading eight as green diamonds, and the remaining systems as a hexbin density layer on a logarithmic color scale. **b**, Zoomed-in scatter plot of the consistent discriminating systems, colored by strategy (PDISC, ODISC, DISC, CDISC). No jitter is applied in either panel, so that small differences between individual systems can be read off reliably. This figure is theunjittered counterpart of Fig. 2 of the main text.

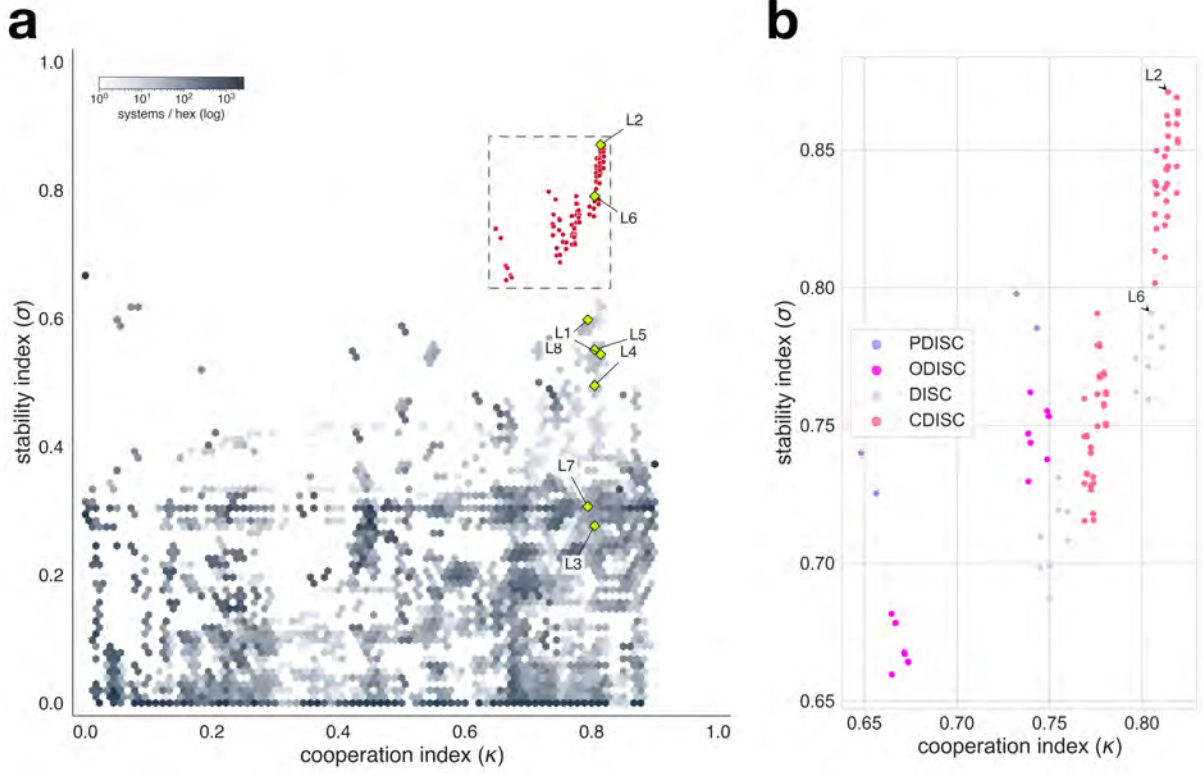


Figure 14: Stability and cooperation indices of all moral systems, without jitter, for $\chi = 0.02$ and $\epsilon = 0.1$. **a**, Each of the 524,800 moral systems that are distinct up to the relabeling of reputations (one representative per mirror class; Section 4.1) is positioned by its resident cooperation index (κ) and average stability index (σ); the consistent discriminating systems are shown as red circles, the leading eight as green diamonds, and the remaining systems as a hexbin density layer on a logarithmic color scale. **b**, Zoomed-in scatter plot of the consistent discriminating systems, colored by strategy (PDISC, ODISC, DISC, CDISC). No jitter is applied in either panel, so that small differences between individual systems can be read off reliably.

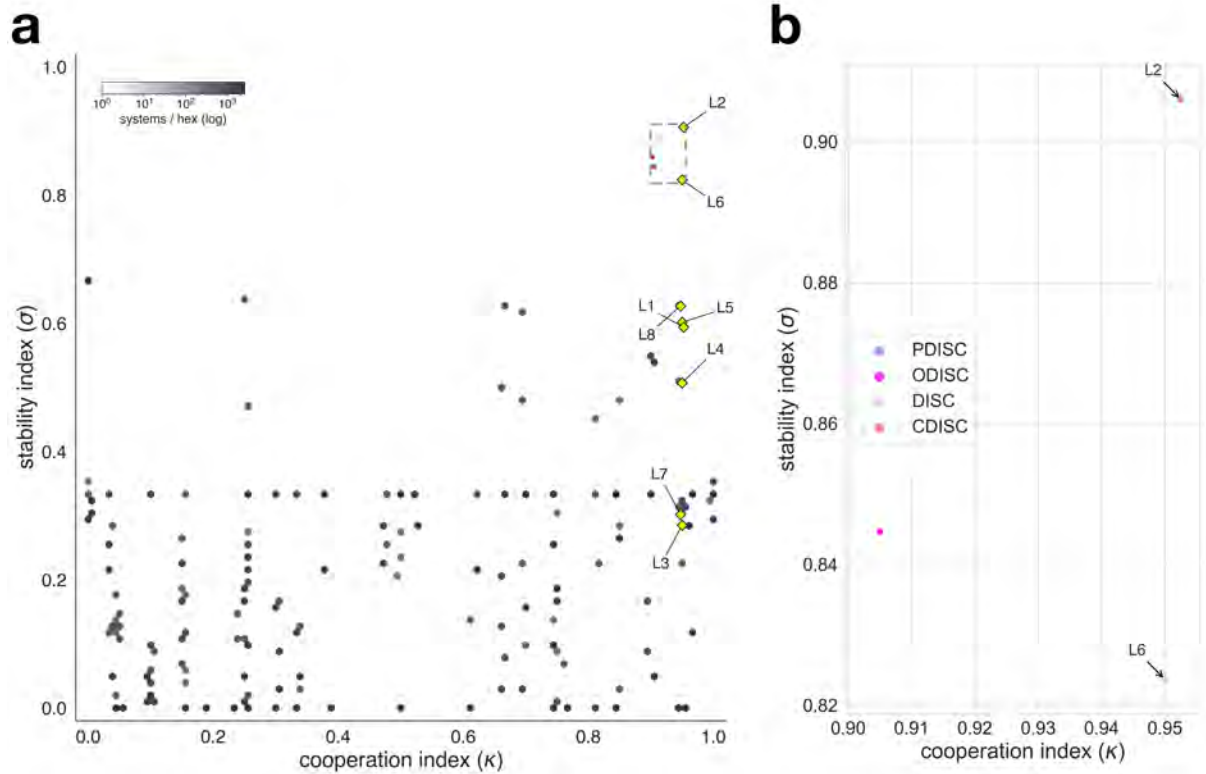


Figure 15: Stability and cooperation indices of all moral systems, without jitter, for $\chi = 0.05$ and $\epsilon = 0$. **a**, Each of the 524,800 moral systems that are distinct up to the relabeling of reputations (one representative per mirror class; Section 4.1) is positioned by its resident cooperation index (κ) and average stability index (σ); the consistent discriminating systems are shown as red circles, the leading eight as green diamonds, and the remaining systems as a hexbin density layer on a logarithmic color scale. **b**, Zoomed-in scatter plot of the consistent discriminating systems, colored by strategy (PDISC, ODISC, DISC, CDISC). No jitter is applied in either panel, so that small differences between individual systems can be read off reliably.

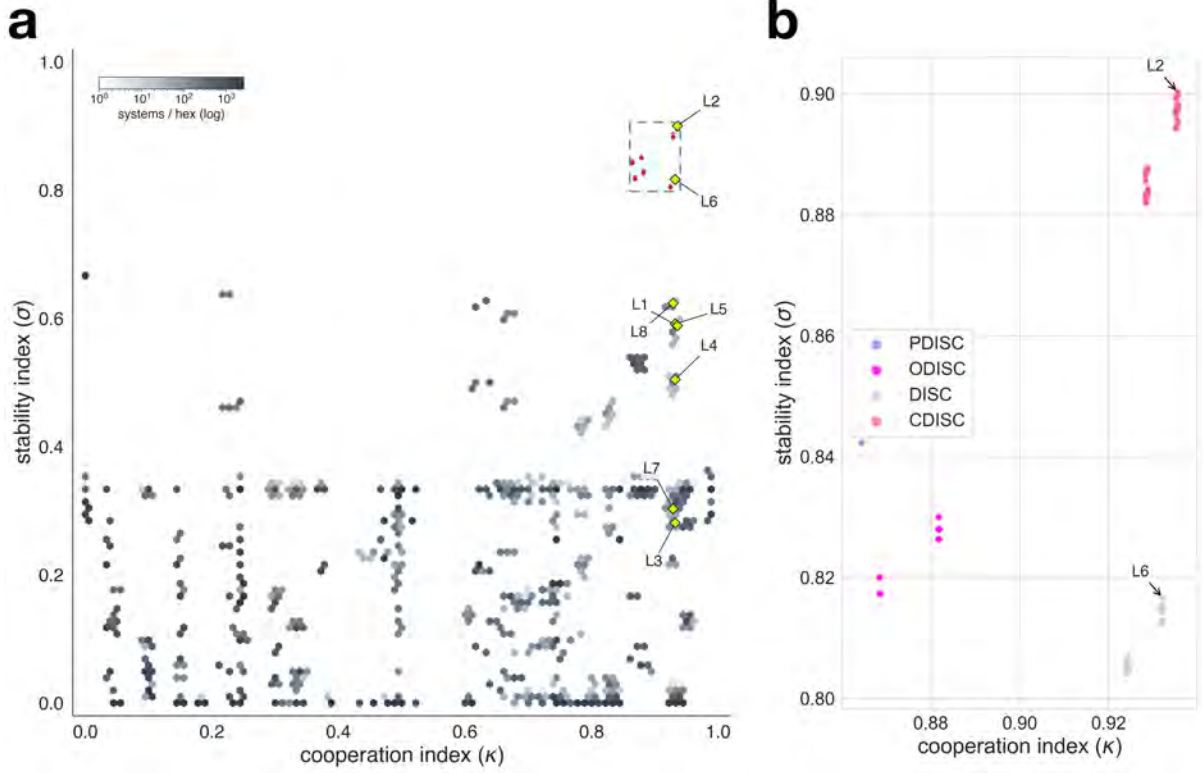


Figure 16: Stability and cooperation indices of all moral systems, without jitter, for $\chi = 0.05$ and $\epsilon = 0.01$. **a**, Each of the 524,800 moral systems that are distinct up to the relabeling of reputations (one representative per mirror class; Section 4.1) is positioned by its resident cooperation index (κ) and average stability index (σ); the consistent discriminating systems are shown as red circles, the leading eight as green diamonds, and the remaining systems as a hexbin density layer on a logarithmic color scale. **b**, Zoomed-in scatter plot of the consistent discriminating systems, colored by strategy (PDISC, ODISC, DISC, CDISC). No jitter is applied in either panel, so that small differences between individual systems can be read off reliably.

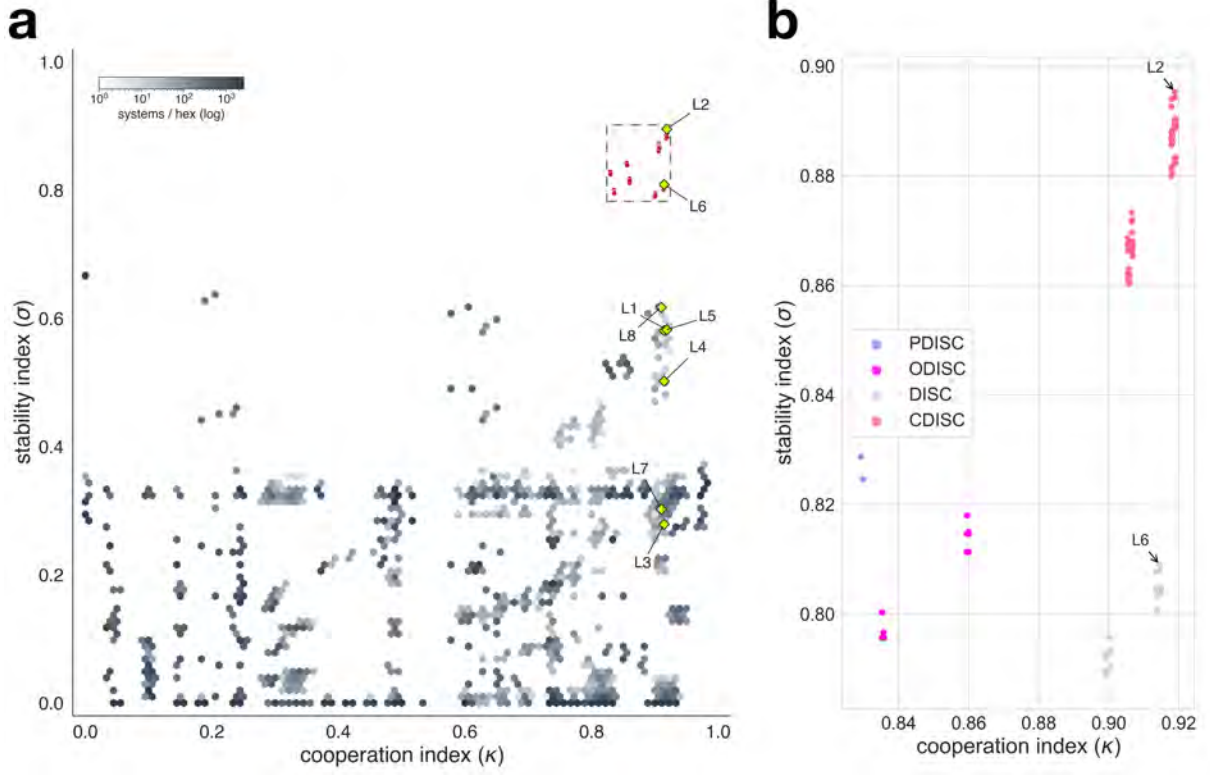


Figure 17: Stability and cooperation indices of all moral systems, without jitter, for $\chi = 0.05$ and $\epsilon = 0.02$. **a**, Each of the 524,800 moral systems that are distinct up to the relabeling of reputations (one representative per mirror class; Section 4.1) is positioned by its resident cooperation index (κ) and average stability index (σ); the consistent discriminating systems are shown as red circles, the leading eight as green diamonds, and the remaining systems as a hexbin density layer on a logarithmic color scale. **b**, Zoomed-in scatter plot of the consistent discriminating systems, colored by strategy (PDISC, ODISC, DISC, CDISC). No jitter is applied in either panel, so that small differences between individual systems can be read off reliably.

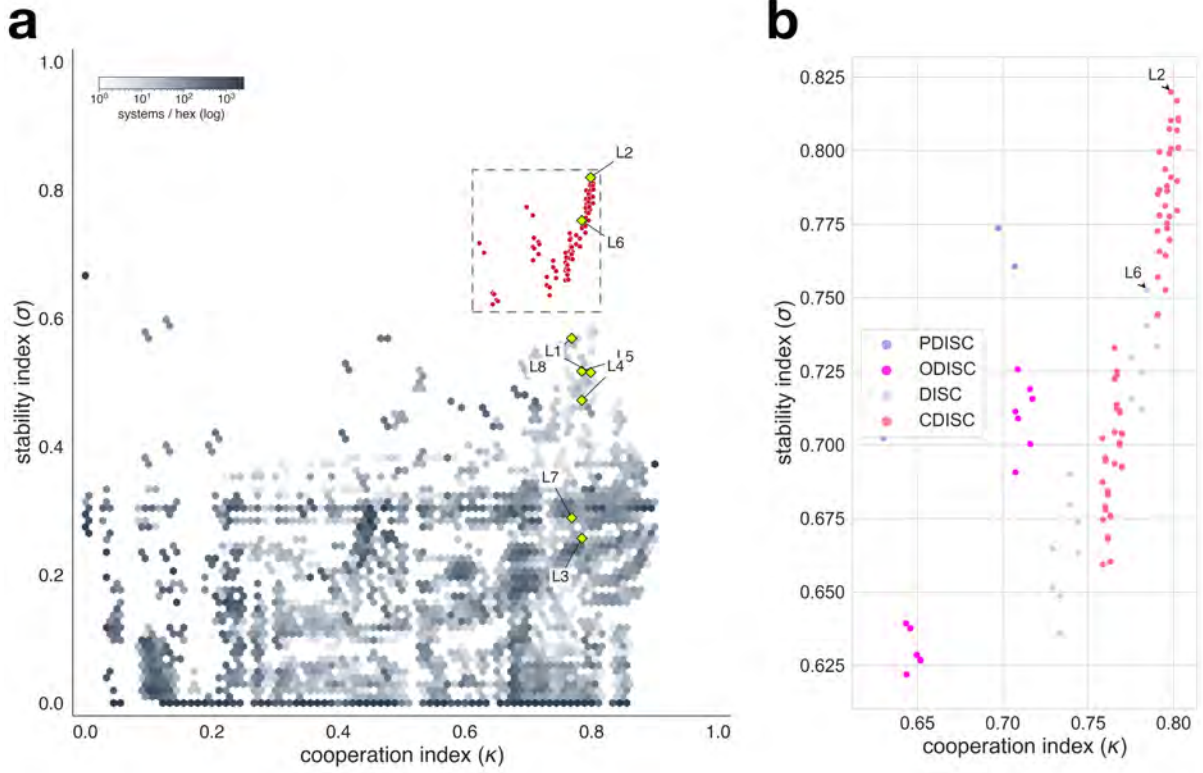


Figure 18: Stability and cooperation indices of all moral systems, without jitter, for $\chi = 0.05$ and $\epsilon = 0.1$. **a**, Each of the 524,800 moral systems that are distinct up to the relabeling of reputations (one representative per mirror class; Section 4.1) is positioned by its resident cooperation index (κ) and average stability index (σ); the consistent discriminating systems are shown as red circles, the leading eight as green diamonds, and the remaining systems as a hexbin density layer on a logarithmic color scale. **b**, Zoomed-in scatter plot of the consistent discriminating systems, colored by strategy (PDISC, ODISC, DISC, CDISC). No jitter is applied in either panel, so that small differences between individual systems can be read off reliably.

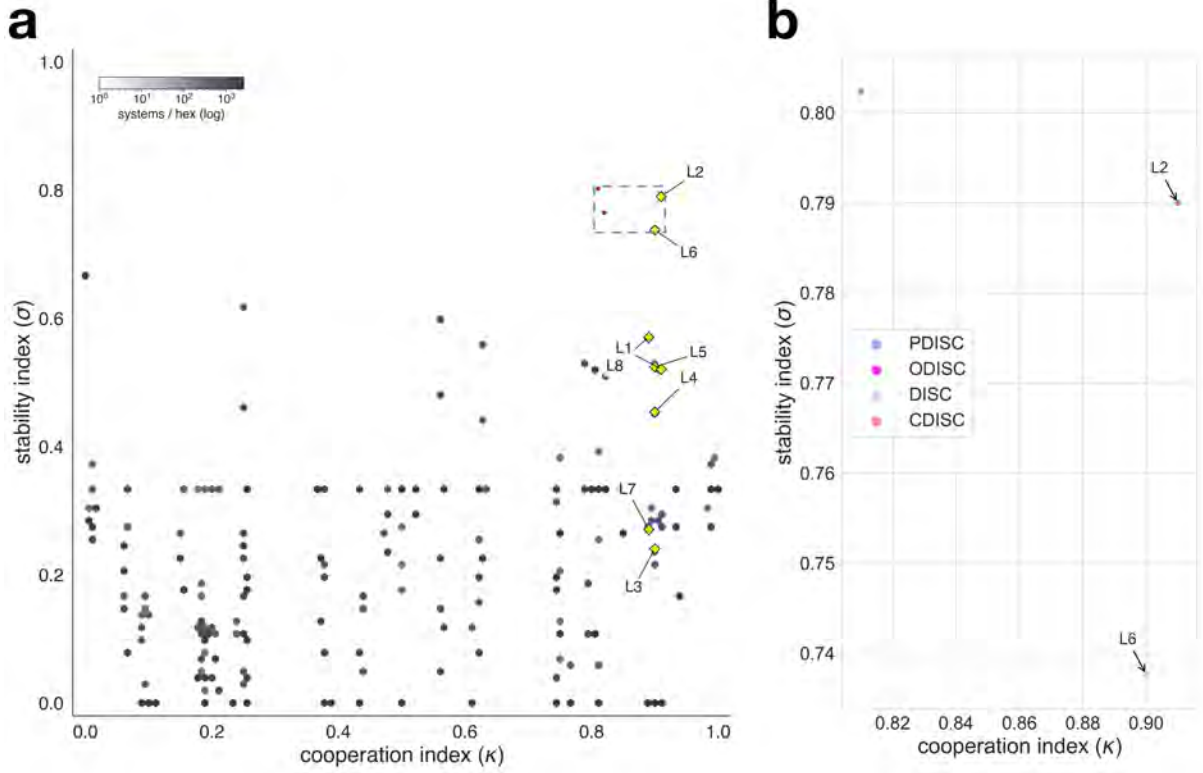


Figure 19: Stability and cooperation indices of all moral systems, without jitter, for $\chi = 0.1$ and $\epsilon = 0$. **a**, Each of the 524,800 moral systems that are distinct up to the relabeling of reputations (one representative per mirror class; Section 4.1) is positioned by its resident cooperation index (κ) and average stability index (σ); the consistent discriminating systems are shown as red circles, the leading eight as green diamonds, and the remaining systems as a hexbin density layer on a logarithmic color scale. **b**, Zoomed-in scatter plot of the consistent discriminating systems, colored by strategy (PDISC, ODISC, DISC, CDISC). No jitter is applied in either panel, so that small differences between individual systems can be read off reliably.

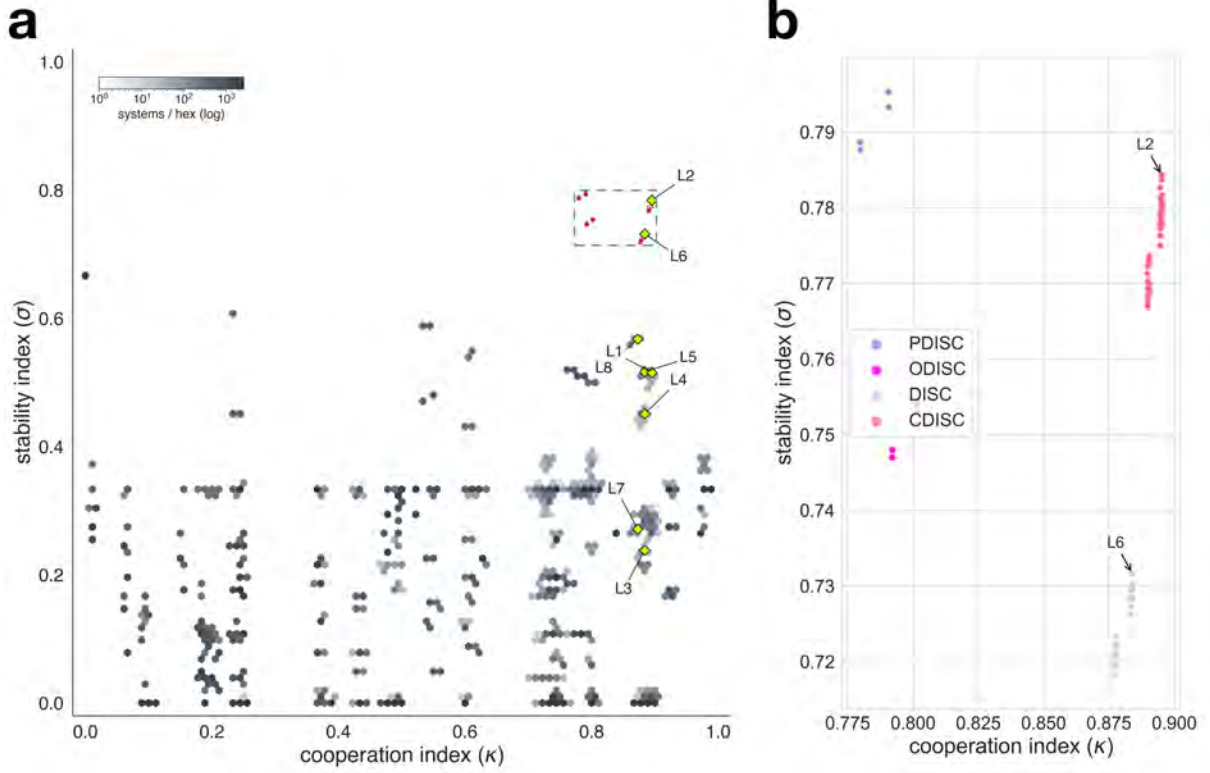


Figure 20: Stability and cooperation indices of all moral systems, without jitter, for $\chi = 0.1$ and $\epsilon = 0.01$. **a**, Each of the 524,800 moral systems that are distinct up to the relabeling of reputations (one representative per mirror class; Section 4.1) is positioned by its resident cooperation index (κ) and average stability index (σ); the consistent discriminating systems are shown as red circles, the leading eight as green diamonds, and the remaining systems as a hexbin density layer on a logarithmic color scale. **b**, Zoomed-in scatter plot of the consistent discriminating systems, colored by strategy (PDISC, ODISC, DISC, CDISC). No jitter is applied in either panel, so that small differences between individual systems can be read off reliably.

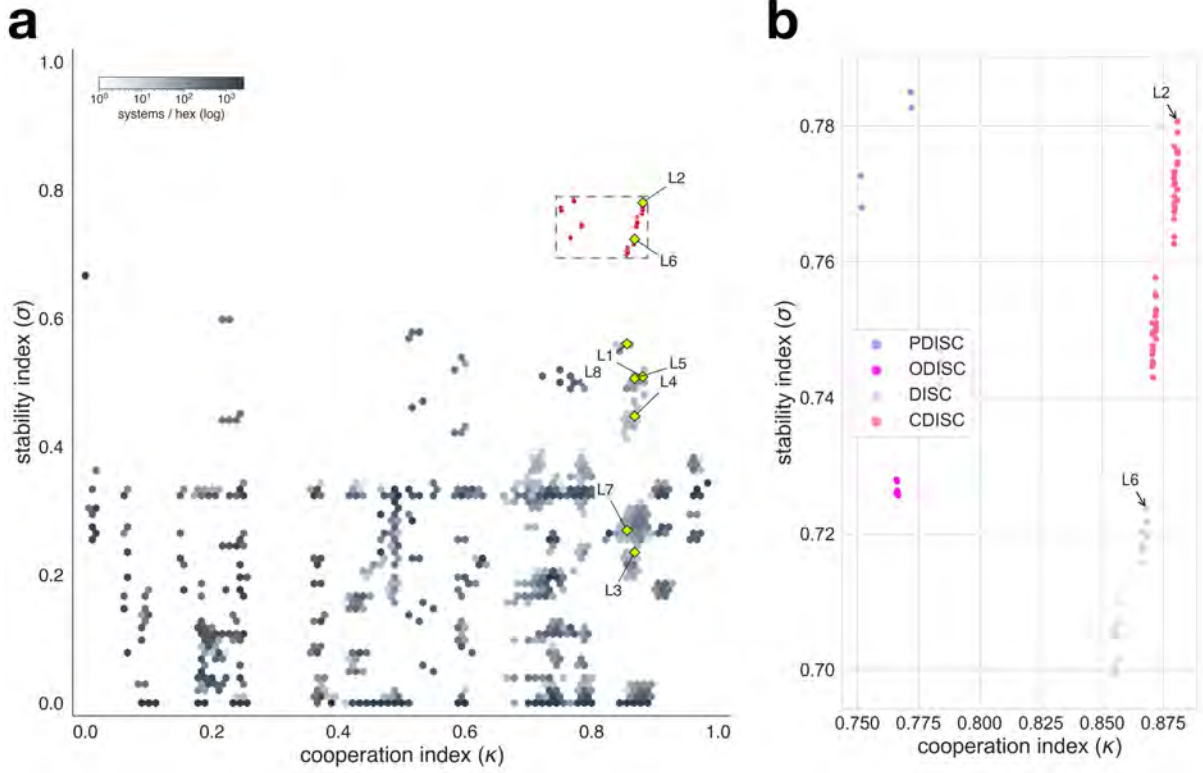


Figure 21: Stability and cooperation indices of all moral systems, without jitter, for $\chi = 0.1$ and $\epsilon = 0.02$. **a**, Each of the 524,800 moral systems that are distinct up to the relabeling of reputations (one representative per mirror class; Section 4.1) is positioned by its resident cooperation index (κ) and average stability index (σ); the consistent discriminating systems are shown as red circles, the leading eight as green diamonds, and the remaining systems as a hexbin density layer on a logarithmic color scale. **b**, Zoomed-in scatter plot of the consistent discriminating systems, colored by strategy (PDISC, ODISC, DISC, CDISC). No jitter is applied in either panel, so that small differences between individual systems can be read off reliably.

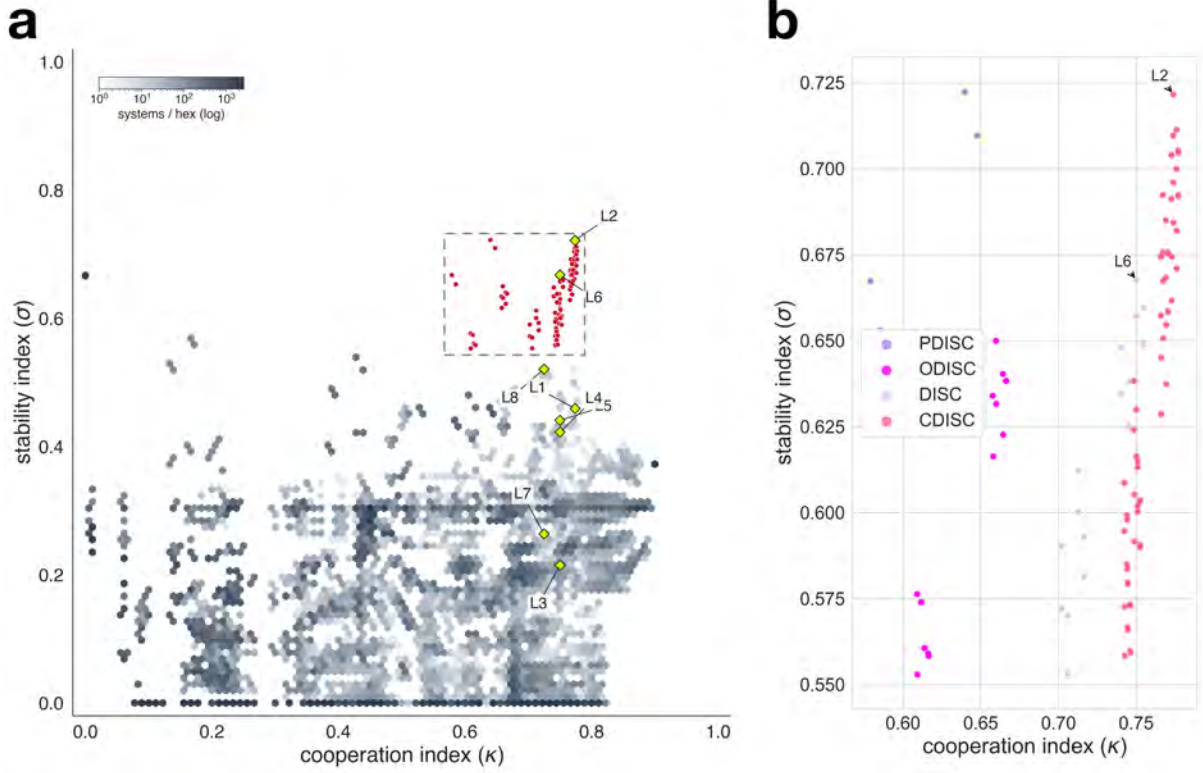


Figure 22: Stability and cooperation indices of all moral systems, without jitter, for $\chi = 0.1$ and $\epsilon = 0.1$. **a**, Each of the 524,800 moral systems that are distinct up to the relabeling of reputations (one representative per mirror class; Section 4.1) is positioned by its resident cooperation index (κ) and average stability index (σ); the consistent discriminating systems are shown as red circles, the leading eight as green diamonds, and the remaining systems as a hexbin density layer on a logarithmic color scale. **b**, Zoomed-in scatter plot of the consistent discriminating systems, colored by strategy (PDISC, ODISC, DISC, CDISC). No jitter is applied in either panel, so that small differences between individual systems can be read off reliably.

4.4 Disaggregated stability indices

The average stability index σ (32) used in Fig. 2 of the main text compresses the three classes of cooperative dilemma into a single number, so that systems with similar averages may owe their stability to very different classes. Figs. 23–25 therefore plot the stability index of each class separately. At the small error rates considered ($\chi = \epsilon = 0.02$), the Monte Carlo estimates lie close to the limiting indices derived analytically in Sections 3.6 and 3.7 (Tables 15 and 22), which they thus corroborate.

The zoomed-in panels b show how the stability of the consistent discriminating systems is distributed across classes. Except for those using CDISC, all consistent discriminating systems have one weak class, in which they fall well below the rest and cluster near the limiting index of $2/3$: the systems using ODISC or PDISC in prisoner’s dilemmas ($\sigma^{\text{PD}} \approx 0.6$; Fig. 23b), and the systems using DISC in snowdrift games ($\sigma^{\text{SG}} \approx 0.7$; Fig. 24b). No system is weak in stag hunts, where all consistent discriminating systems are highly stable ($\sigma^{\text{SH}} > 0.92$; Fig. 25b). Within any single class, the systems using CDISC need not be the most stable: those using DISC attain comparable values in prisoner’s dilemmas, and those using ODISC or PDISC are stable in essentially every stag hunt ($\sigma^{\text{SH}} \approx 1$). What distinguishes the systems using CDISC is instead the absence of a weak class, in line with their universal stability (Theorem 1).

The disaggregated indices also clarify the position of the leading eight (green diamonds in panels a). All eight are highly stable in prisoner’s dilemmas ($\sigma^{\text{PD}} \approx 0.9$; Fig. 23a), the class that contains the donation game for which they were originally derived; outside it, they scatter in line with Table 22. In snowdrift games (Fig. 24a), only L1 and L2 remain highly stable; L4 and L6 sit near their limiting index of $2/3$, alongside the consistent systems using DISC, while L3, L5, L7, and L8, which can be invaded by CDISC in every snowdrift game, drop to $\sigma^{\text{SG}} \approx 0$. In stag hunts (Fig. 25a), L2, L5, L6, and L8 remain highly stable, while L1, L3, L4, and L7, which can be invaded by unconditional cooperators, fall to $\sigma^{\text{SH}} \approx 0$. L2 is thus the only leading-eight system that is highly stable in all three classes, and the high average stability of L6 in Fig. 2 of the main text rests on prisoner’s dilemmas and stag hunts alone.

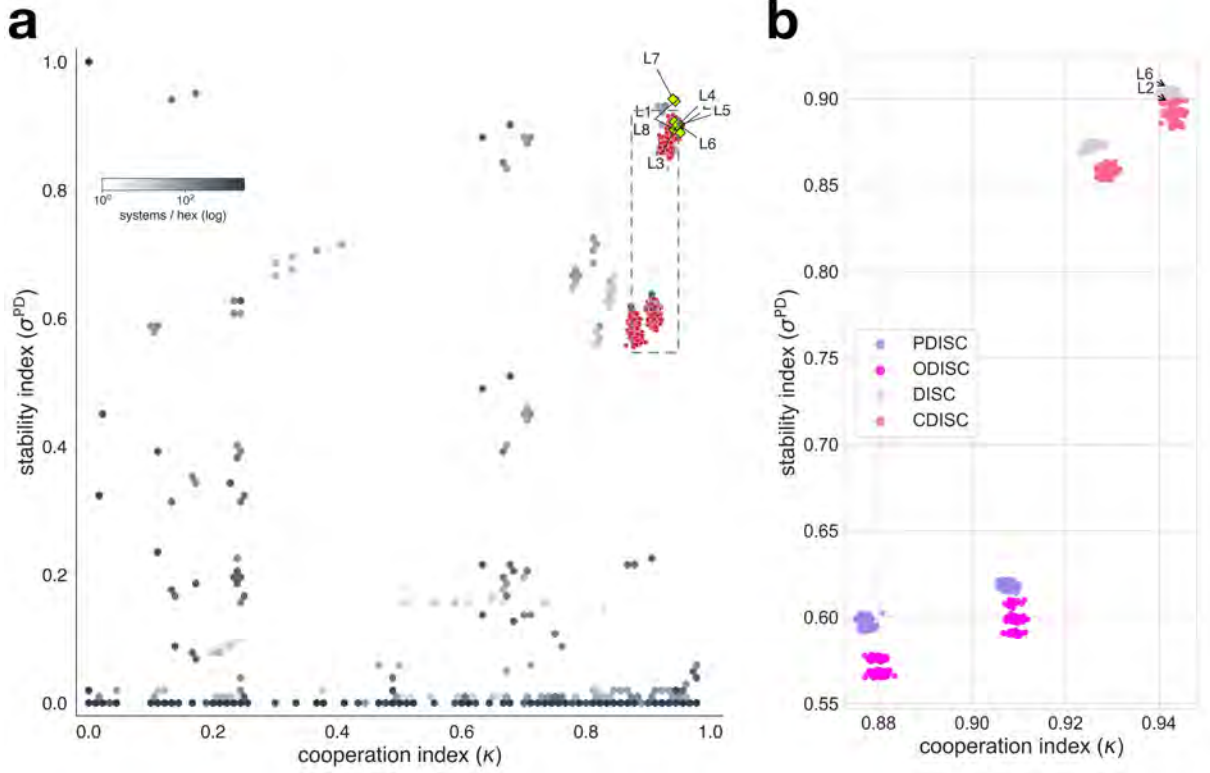


Figure 23: Stability and cooperation indices of all moral systems in the space of prisoner's dilemmas only, for $\chi = \epsilon = 0.02$. Unlike Fig. 2 of the main text, the vertical axis is the stability index of a single class of dilemma, σ^{PD} , rather than the average σ across the three classes. **a**, Each of the 524,800 moral systems that are distinct up to the relabeling of reputations (one representative per mirror class; Section 4.1) is positioned by its resident cooperation index (κ) and its prisoner's dilemma stability index (σ^{PD}); the consistent discriminating systems are shown as red circles, the leading eight as green diamonds, and the remaining systems as a hexbin density layer on a logarithmic color scale. **b**, Zoomed-in scatter plot of the consistent discriminating systems, colored by strategy (PDISC, ODISC, DISC, CDISC). Points are jittered for clarity in both panels.

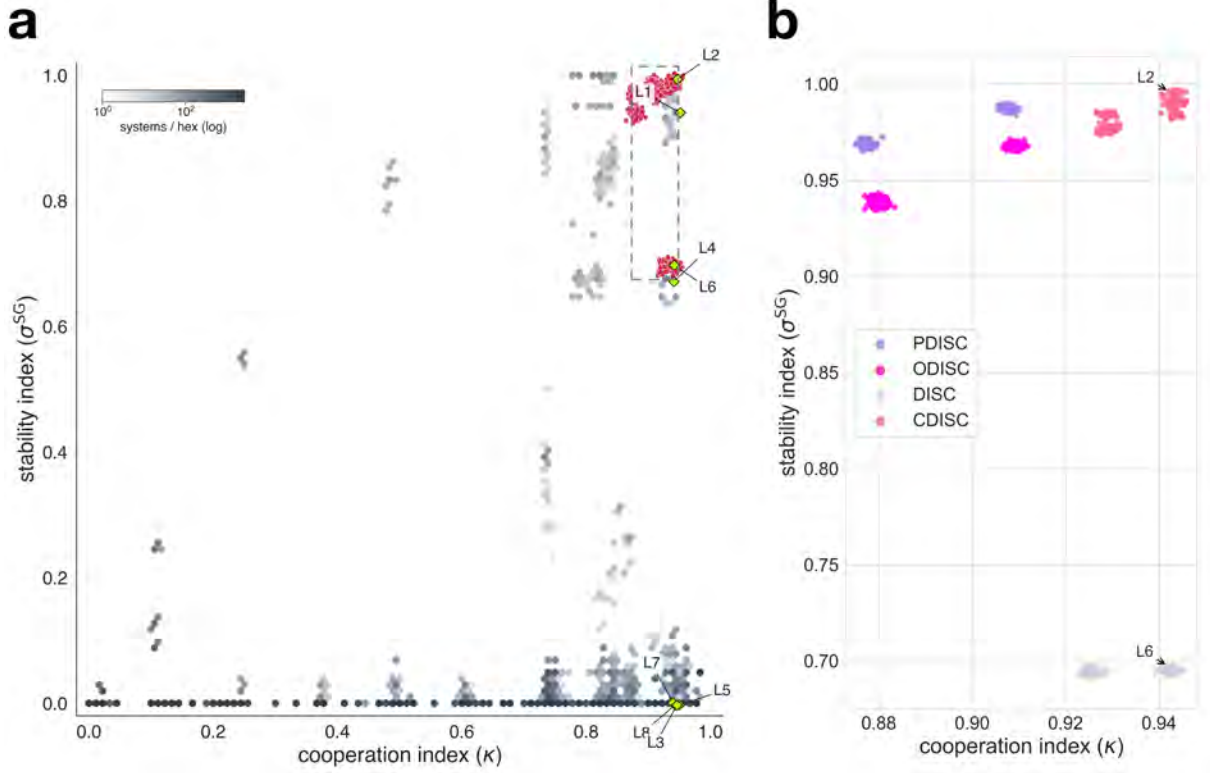


Figure 24: Stability and cooperation indices of all moral systems in the space of snowdrift games only, for $\chi = \epsilon = 0.02$. Unlike Fig. 2 of the main text, the vertical axis is the stability index of a single class of dilemma, σ^{SG} , rather than the average σ across the three classes. **a**, Each of the 524,800 moral systems that are distinct up to the relabeling of reputations (one representative per mirror class; Section 4.1) is positioned by its resident cooperation index (κ) and its snowdrift game stability index (σ^{SG}); the consistent discriminating systems are shown as red circles, the leading eight as green diamonds, and the remaining systems as a hexbin density layer on a logarithmic color scale. **b**, Zoomed-in scatter plot of the consistent discriminating systems, colored by strategy (PDISC, ODISC, DISC, CDISC). Points are jittered for clarity in both panels.

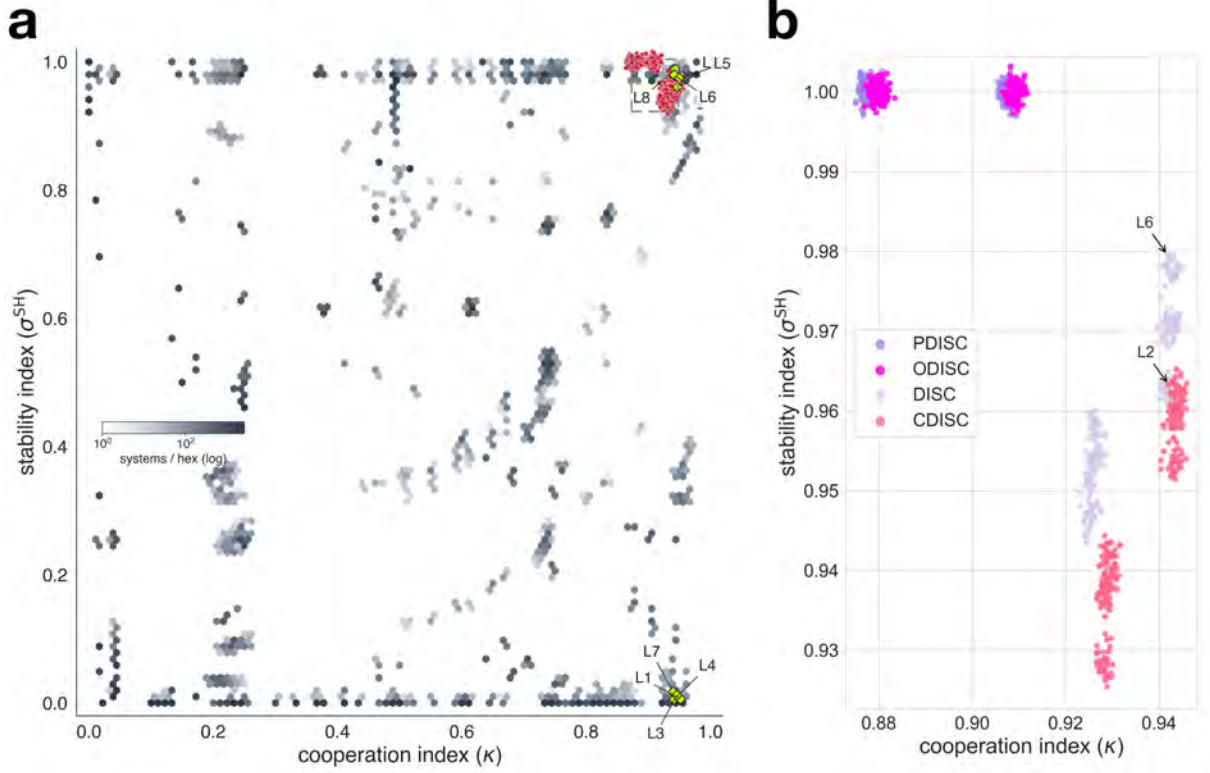


Figure 25: Stability and cooperation indices of all moral systems in the space of stag hunts only, for $\chi = \epsilon = 0.02$. Unlike Fig. 2 of the main text, the vertical axis is the stability index of a single class of dilemma, σ^{SH} , rather than the average σ across the three classes. **a**, Each of the 524,800 moral systems that are distinct up to the relabeling of reputations (one representative per mirror class; Section 4.1) is positioned by its resident cooperation index (κ) and its stag hunt stability index (σ^{SH}); the consistent discriminating systems are shown as red circles, the leading eight as green diamonds, and the remaining systems as a hexbin density layer on a logarithmic color scale. **b**, Zoomed-in scatter plot of the consistent discriminating systems, colored by strategy (PDISC, ODISC, DISC, CDISC); note the compressed vertical scale, spanning less than 0.08. Points are jittered for clarity in both panels.

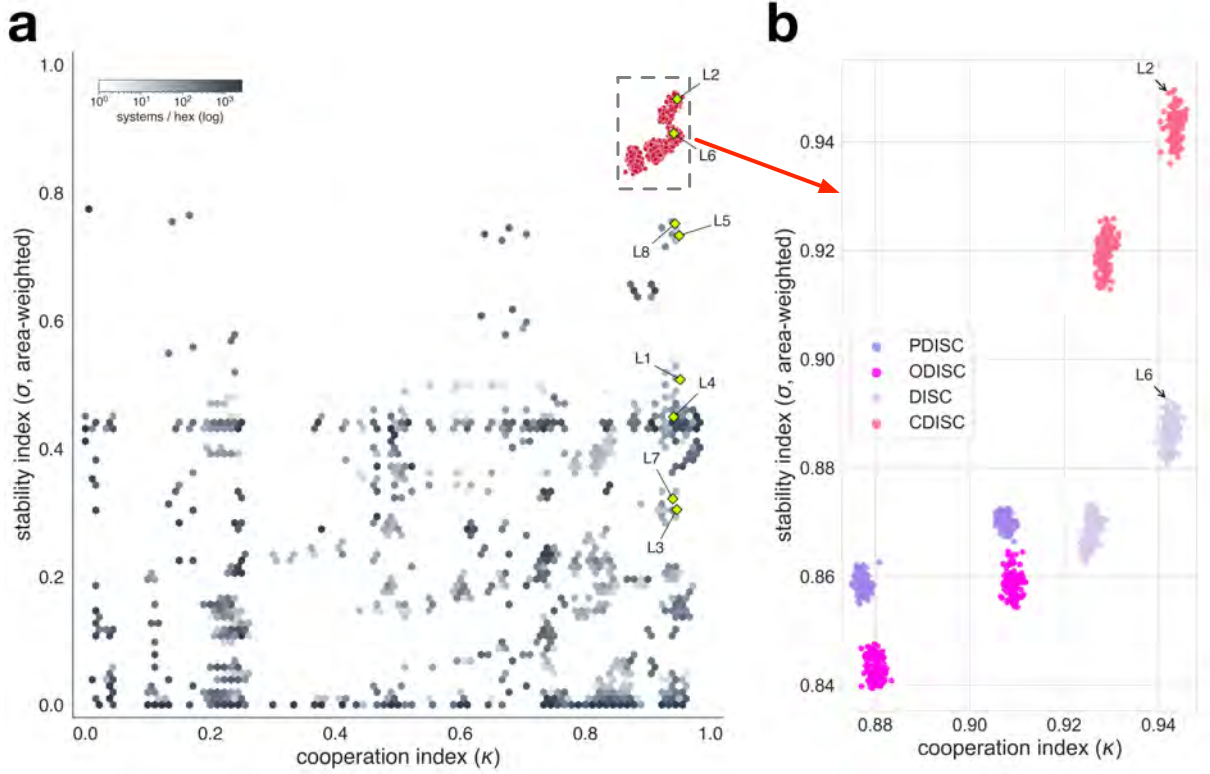


Figure 26: Stability and cooperation indices of all moral systems under the area-weighted index σ^w of Section 1.10.3, for $\chi = \epsilon = 0.02$; compare Fig. 2 of the main text, which uses the equally weighted index σ . **a**, Each of the 524,800 moral systems that are distinct up to the relabeling of reputations (one representative per mirror class; Section 4.1) is positioned by its resident cooperation index (κ) and area-weighted stability index (σ^w); the consistent discriminating systems are shown as red circles, the leading eight as green diamonds, and the remaining systems as a hexbin density layer on a logarithmic color scale. **b**, Zoomed-in scatter plot of the consistent discriminating systems (jittered for clarity), colored by strategy (PDISC, ODISC, DISC, CDISC); as in Fig. 2b, the systems using CDISC achieve the highest stability.

4.5 Area-weighted stability index

The average stability index σ (32) gives the three classes of cooperative dilemma equal weight. The area-weighted index σ^w of Section 1.10.3 instead weights each class in proportion to the area of its normalized payoff region. Figs. 26 and 27 show that this choice makes no qualitative difference. The consistent discriminating systems remain the only cluster to combine high stability with high cooperation, both at the baseline error rates – where the systems using CDISC are again the most stable of all – and across the error-rate grid of Fig. 5. The main visible change is a tighter gap between the cluster and the rest, as the ceiling for systems outside the cluster rises from $\sigma = 2/3$ to $\sigma^w = 7/9$ (exact values, attained at every error-rate combination we examine); both are attained by fully uncooperative systems ($\kappa = 0$), which are stable in every prisoner's dilemma and stag hunt but in no snowdrift game.

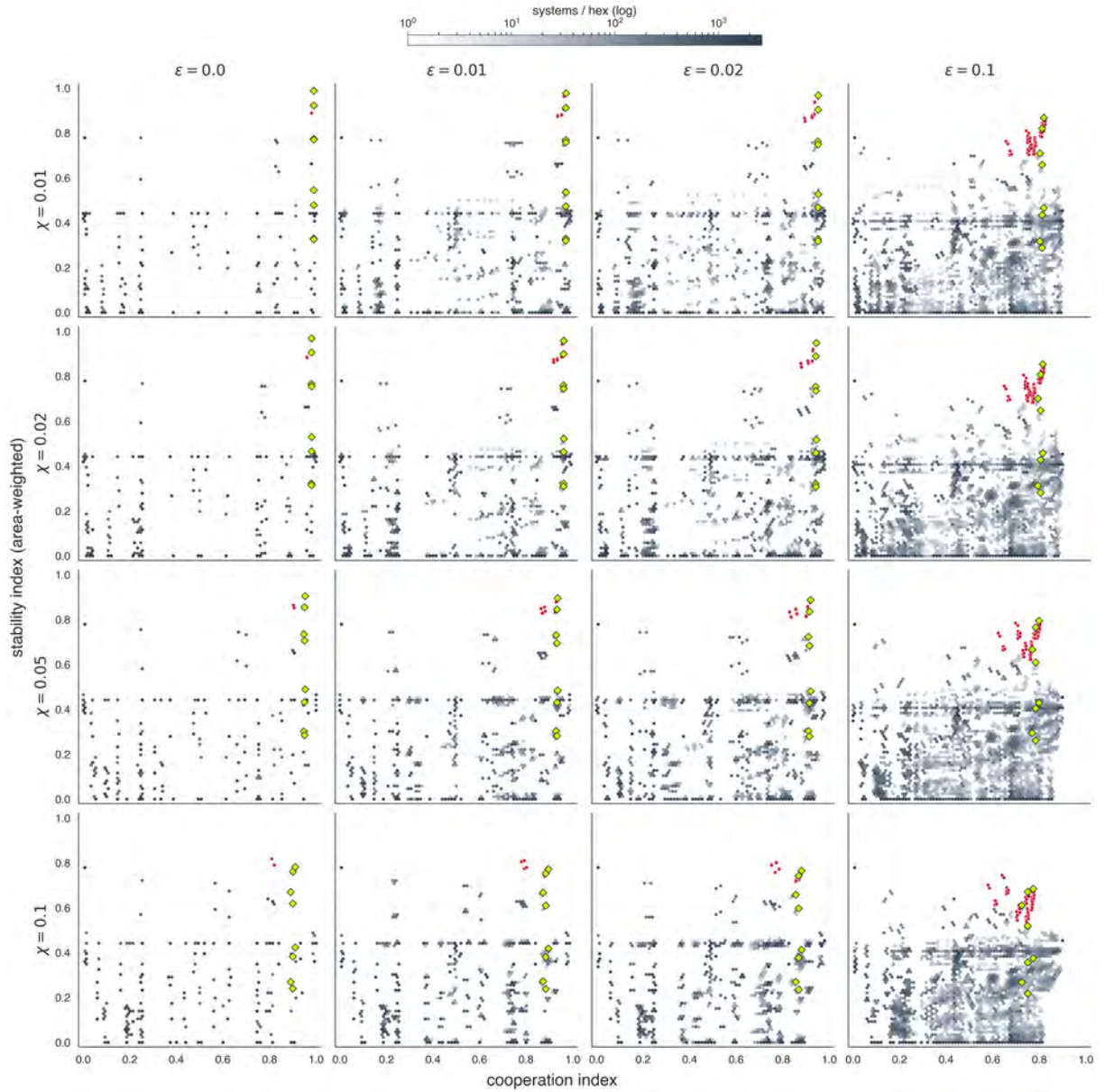


Figure 27: Stability and cooperation indices across a range of error rates, under the area-weighted index σ^w of Section 1.10.3; compare Fig. 5, which uses the equally weighted index σ . Each panel shows, for one combination of execution error rate $\epsilon \in \{0, 0.01, 0.02, 0.1\}$ (columns) and assignment error rate $\chi \in \{0.01, 0.02, 0.05, 0.1\}$ (rows), the cooperation index (κ) and area-weighted stability index (σ^w) of each of the 524,800 mirror-class representatives (Section 4.1). The consistent discriminating systems are shown as red circles, the leading eight as green diamonds, and the remaining systems as a hexbin density layer on a logarithmic color scale.

References

1. Peña, J. & Nöldeke, G. Cooperative dilemmas with binary actions and multiple players. *Dynamic Games and Applications* **13**, 1156–1193 (2023).
2. Ohtsuki, H. & Iwasa, Y. How should we define goodness?—reputation dynamics in indirect reciprocity. *Journal of Theoretical Biology* **231**, 107–20 (2004).
3. Nakamura, M. & Ohtsuki, H. Indirect reciprocity in three types of social dilemmas. *Journal of Theoretical Biology* **355**, 117–127 (2014).
4. Ohtsuki, H. & Iwasa, Y. The leading eight: Social norms that can maintain cooperation by indirect reciprocity. *Journal of Theoretical Biology* **239**, 435–444 (2006).
5. van Lint, J. H. & Wilson, R. M. *A Course in Combinatorics* (Cambridge University Press, Cambridge, 2001), 2nd edn.