Article

# The evolution of private reputations in information-abundant landscapes

https://doi.org/10.1038/s41586-024-07977-x

Received: 23 June 2023

Accepted: 21 August 2024

Published online: 25 September 2024

Check for updates

Sebastián Michel-Mata[1], Mari Kawakatsu[2,3], Joseph Sartini[4,5], Taylor A. Kessinger[2], Joshua B. Plotkin[2,3] & Corina E. Tarnita[1]✉

Reputations are critical to human societies, as individuals are treated differently based on their social standing[1,2]. For instance, those who garner a good reputation by helping others are more likely to be rewarded by third parties[3–5]. Achieving widespread cooperation in this way requires that reputations accurately reflect behaviour[6] and that individuals agree about each other's standings[7]. With few exceptions[8–10], theoretical work has assumed that information is limited, which hinders consensus[7,11] unless there are mechanisms to enforce agreement, such as empathy[12], gossip[13–15] or public institutions[16]. Such mechanisms face challenges in a world where empathy, effective communication and institutional trust are compromised[17–19]. However, information about others is now abundant and readily available, particularly through social media. Here we demonstrate that assigning private reputations by aggregating several observations of an individual can accurately capture behaviour, foster emergent agreement without enforcement mechanisms and maintain cooperation, provided individuals exhibit some tolerance for bad actions. This finding holds for both first- and second-order norms of judgement and is robust even when norms vary within a population. When the aggregation rule itself can evolve, selection indeed favours the use of several observations and tolerant judgements. Nonetheless, even when information is freely accessible, individuals do not typically evolve to use all of it. This method of assessing reputations—'look twice, forgive once', in a nutshell—is simple enough to have arisen early in human culture and powerful enough to persist as a fundamental component of social heuristics.

The theory of indirect reciprocity helps explain how cooperation can arise when individuals lack a shared history of social interactions[3,20,21]. Without personal experience, we often decide how to interact with someone based on our information about their past interactions with others[22]. In the simplest model, everyone engages in pairwise interactions through a one-shot donation game (a simplified prisoner's dilemma) between a potential donor and a recipient. The donor can either cooperate, paying a cost to provide a benefit to the recipient, or defect, incurring no cost and generating no benefit. Donors can act unconditionally by always cooperating (ALLC) or always defecting (ALLD), or they can condition their behaviour on the recipient's reputation (discriminate or DISC), donating only if they perceive the recipient to have good social standing. The donor's action may result in the donor gaining a good reputation in the eyes of a third party, so that the donor may expect to receive donations from the third party, hence the term indirect reciprocity.

In most models, reputations are binary (good or bad)[4,23], and they are assigned to an individual by observing and assessing a single previous action (out of many), an assumption we refer to as 'limited information'. To assess whether an action is good or bad, observers use a shared social norm. The simplest norm is called scoring[24–26]. This first-order norm uses only the donor's action to assign a reputation, assessing cooperative actions as good and defections as bad. However, scoring has a weakness that eventually leads a population to defection[4,8,23]: discriminators will choose to defect against defectors, who have bad reputations, but in so doing, they receive bad reputations themselves. In contrast, unconditional cooperators pay no such reputational cost, so they outcompete discriminators. Once unconditional cooperators are common in the population, they will, in turn, be outcompeted by unconditional defectors. Thus, DISC can be invaded by ALLC, which opens the door for ALLD to invade (Fig. 1a). This weakness is called the scoring dilemma[27], and it can be addressed using higher-order social norms[28].

Higher-order norms incorporate more information to help distinguish between justified and unjustified acts of defection[29]. For example, second-order norms—such as the three we consider: stern-judging, simple-standing and shunning (see Methods for definitions)—assess a donor using the donor's action and the recipient's reputation. Thus, defection may sometimes be judged positively[30]: for example, under stern-judging, someone who defects against a recipient in bad standing

[1]Department of Ecology and Evolutionary Biology, Princeton University, Princeton, NJ, USA. [2]Department of Biology, University of Pennsylvania, Philadelphia, PA, USA. [3]Center for Mathematical Biology, University of Pennsylvania, Philadelphia, PA, USA. [4]Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ, USA. [5]Present address: Department of Biostatistics, Johns Hopkins University, Baltimore, MD, USA. ✉e-mail: ctarnita@princeton.edu

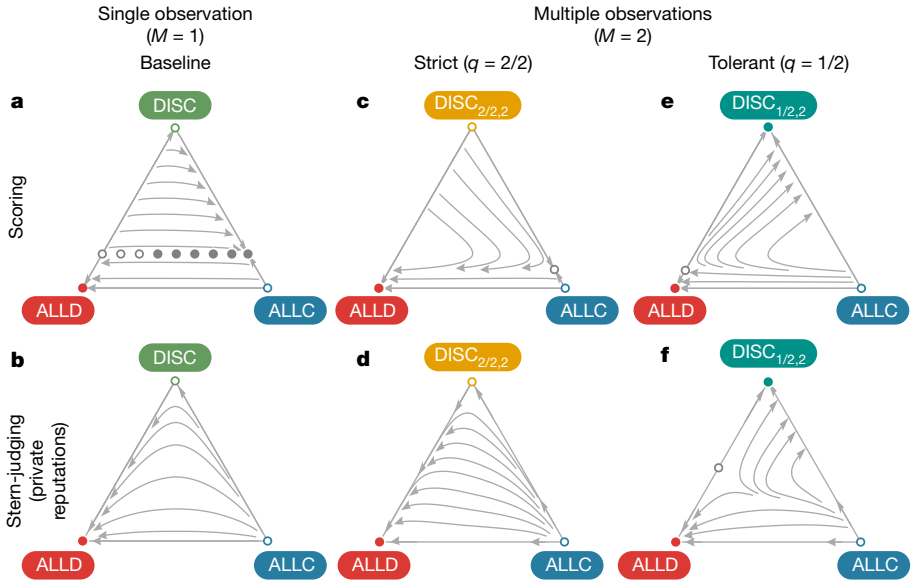



**Fig. 1 | 'Look twice, forgive once' solves both the scoring and the punishment dilemmas. a**–**f**, Simplexes show the dynamics of competition among the strategies ALLC, ALLD and a discriminator strategy under scoring (**a**,**c**,**e**) and stern-judging (**b**,**d**,**f**). Arrows represent evolutionary dynamics. Filled (empty) points represent stable (unstable) equilibria. Discriminators use either a single observation (DISC; **a**,**b**) or several observations ($DISC_{q,M}$; **c**–**f**) to assign reputations. **a**, DISC is stable only in mixed equilibria that include some ALLC. **b**, ALLD dominates (the only stable equilibrium). **c**,**d**, ALLD dominates. **e**, 'Look twice, forgive once' ($DISC_{1/2,2}$) outcompetes ALLC and is stable, with the largest basin of attraction. At this equilibrium, the rate of cooperation (the fraction of cooperative interactions) is over 99.8%. **f**, 'Look twice, forgive once' ($DISC_{1/2,2}$) outcompetes ALLD when common and is stable, with a substantial basin of attraction. At this equilibrium, the rate of cooperation is over 99.8%. For all panels, the benefit-to-cost ratio was $b/c = 5$, with error rates of assessment and execution $\alpha = 0.02$ and $\varepsilon = 0.02$, respectively. Corresponding results for simple-standing and shunning are shown in Extended Data Fig. 1.

earns a good reputation[31]. Although higher-order norms avoid the scoring dilemma[6,32], they have a significant drawback: a donor's defection will appear justified to an observer only when the observer and the donor agree on the recipient's reputation. Thus, under higher-order norms, cooperation depends on population-level agreement about an individual's reputation. Agreement is ensured when reputations are broadcast publicly, for example, by an institution[16] or through rapid gossip[13–15]. However, when each observer forms an independent opinion of every donor, so that reputations are held privately, two individuals may disagree about the reputation of a third simply because they observed the third individual in different interactions (for example, if X observes Z interacting with A whereas Y observes Z interacting with B, then X and Y may form different opinions of Z). This weakness of private reputations is further exacerbated by errors in either the execution of the strategy[4,23] or the assessment of the action[33,34], which can cause disagreement even when two observers judge the same interaction. In the absence of agreement-enforcing mechanisms, disagreement propagates and leads people to judge each other harshly, eventually causing cooperation to collapse[7,11,34,35] (Fig. 1b). This weakness of higher-order norms—that they can sustain cooperation only when people agree about each other's reputations—is called the punishment dilemma[27].

Although agreement-enforcing mechanisms exist, they are vulnerable when empathy, communication and trust in institutions are compromised[12,16–19]. This vulnerability adds to the two dilemmas above to paint a bleak picture for cooperation maintained by reputations. Fortunately, the assumption that reputations are based on limited information is decidedly unrealistic nowadays, when ample information about others is freely and publicly available, especially through social media. Although social interactions are public, the assignment of reputations remains private, as observers can (and do) differ in how they use and evaluate all this information. So, it is both timely and urgent to ask how predictions change when individuals can make more than one observation.

The simple fact that an observer can make several observations of a focal individual's behaviour before forming a judgement should produce more informative reputations: actions that might appear bad if considered in isolation (for example, a mistake) can be assessed in the context of the actor's broader behavioural pattern, so that one bad action will not automatically yield a bad reputation. Previous studies have shown that when several observations are available under the first-order norm scoring, it is possible to find evolutionarily stable discriminators or mixtures of two discriminators[8,9]. In other words, forming judgements from several observations could solve the scoring dilemma. However, it remains unclear whether stable discriminators will be favoured in a dynamic, evolutionary setting or whether individuals will actually evolve to use the information, even when it is freely available. Moreover, it remains unexplored whether using more than one observation to judge others' behaviour can have wider-ranging consequences, for example, by resolving the punishment dilemma under higher-order social norms.

Here we investigate the evolution and social consequences of forming judgements from several observations. As in previous studies[8,9,16], each observer makes $M > 1$ observations of a given donor before assigning them a reputation. To assign a binary reputation based on several actions, each of which is assessed as good or bad according to the observer's social norm, the observer must aggregate such information. We define an aggregation rule as a pair $(q, M)$, where $M$ is the number of observed actions and $q$ is the minimum proportion of those actions that must be good to assign a good reputation to the donor, an approach reminiscent of direct reciprocity strategies like tit-for-$m$-tats[36]. The parameter $q$, thus, serves as a strictness threshold in the aggregation process. For example, $q = 1/M$ corresponds to the least strict (or most tolerant) discriminator because a single good action out of the $M$ observations is sufficient to assign a good reputation. Conversely, $q = M/M$ corresponds to the strictest discriminator because a good reputation is awarded only when all observed actions are assessed as good. We denote as $DISC_{q,M}$ the aggregating discriminator strategy

that uses the $(q, M)$ aggregation rule. $DISC_{1,1}$ is the 'classic' discriminator strategy[4], which relies on a single observation. Additionally, we allow for errors in strategy execution and in the assessment of each observed action[33,34,37]. The dynamics are governed by a payoff-biased imitation process[38,39].

We first investigate the simplest case of $M = 2$ observations and explore whether an aggregating discriminator fares better than the classical, single-observation discriminator against unconditional strategies when the whole population subscribes to the same, fixed social norm. When aggregation is strict, one bad action is enough to assign a bad reputation; so, intuitively, we expect that the strict discriminator ($DISC_{2/2,2}$) will not perform better than classical discriminators. In fact, regardless of the norm (and of the number of observations $M \geq 2$), strict discriminators fare worse against unconditional strategies than classical discriminators do, because observing more interactions increases an observer's chance of seeing an isolated bad incident (Fig. 1c,d and Extended Data Fig. 1). However, if individuals show some tolerance, then aggregating several observations could change outcomes in competition against unconditional strategies. Under the scoring norm, we confirm that such 'tolerant discriminators'[8] ($DISC_{1/2,2}$) are evolutionarily stable against unconditional strategies and, furthermore, that they have a large basin of attraction (Fig. 1e). This is because a discriminator who looks twice and forgives once will overlook some bad actions and can, therefore, view other discriminators as good and cooperate with them. 'Look twice, forgive once' can thereby build a reputation that is good enough to receive cooperation while maintaining the ability to defect against (punish) someone with a bad reputation, which protects it against invasion by unconditional cooperators and resolves the scoring dilemma.

In addition to rescuing reputations under this first-order norm, 'look twice, forgive once' can also maintain cooperative outcomes under higher-order norms, even when reputations are held privately without any agreement-enforcing mechanisms. For example, under stern-judging with private reputations, 'look twice, forgive once' is stable and has a substantial basin of attraction when competing against unconditional strategies (Fig. 1f). This is because tolerant discriminators can find some common ground and cooperate with each other when they agree on the reputation of at least half of the recipients. Such agreement does not help when the population consists mainly of ALLD, as a fraction of them end up with good reputations and receive donations without reciprocating. So, a rare 'look twice, forgive once' mutant cannot invade ALLD. However, when ALLD is rare, discriminators find more good individuals to cooperate with and, by looking twice and forgiving once, they improve their own reputations; it then becomes easier to distinguish the unjustified defections of ALLD and reach a stable equilibrium.

In summary, 'look twice, forgive once' is a stable evolutionary outcome with a substantial basin of attraction for both first-order and second-order social norms, which solves both the scoring and punishment dilemmas. The size of the basin of attraction depends on the norm (Extended Data Fig. 1). However, regardless of the norm, when the population is at the 'look twice, forgive once' equilibrium, a large fraction of interactions is cooperative, so that the equilibrium sustains a high rate of cooperation. Thus, using $M = 2$ observations and being tolerant can promote cooperation across all social norms considered.

Although these results are promising, allowing $M > 1$ observations creates a larger strategy space comprising $M$ possible aggregating discriminators. We must, therefore, evaluate the evolution of discriminating strategies in the context of this full strategy space. To do this for $M = 2$, we consider the dynamics in the three-simplex (tetrahedron) whose vertices correspond to the four strategies ALLC, ALLD, 'look twice, forgive once' ($DISC_{1/2,2}$) and the strict discriminator ($DISC_{2/2,2}$). We find that, regardless of the norm, there is still a substantial basin of attraction towards aggregating discriminators (Fig. 2 and Extended Data Fig. 2) for a wide range of benefit-to-cost ratios (for example, $b/c = 1.5$ to 10), if error rates are not too high (Extended Data Figs. 3 and 4). In the full strategy space, the stable equilibrium can sometimes be a mixture of types, with a majority 'look twice, forgive once' coexisting with a small fraction of strict discriminators (scoring in Fig. 2 and simple-standing in Extended Data Fig. 2). When this happens, the presence of strict strategies at equilibrium may seem like a disadvantage, as it decreases the basin of attraction towards aggregating discriminators under all norms except simple-standing (Extended Data Fig. 2). However, such coexistence has important implications for the robustness of aggregating discriminators. For example, classical DISC and pure aggregating discriminators can be invaded by a strategy that cooperates unconditionally with probability $p$ and defects otherwise. However, a mixture of aggregating discriminators can be stable against invasion by such strategies (Extended Data Fig. 5).

Next, we investigate the evolution of aggregating discriminators in the full strategy space of all discriminating strategies, for fixed $M > 2$, along with the two unconditional strategies. As before, all individuals follow the same, exogenously fixed social norm. To simplify the narrative, henceforth we focus on two norms: the first-order norm, scoring, which is the simplest norm, and one second-order norm, stern-judging, which sustains cooperation when reputations are public but fails when they are private. We find that aggregating discriminators maintain a basin of attraction against unconditional strategies regardless of the number of observations (Fig. 3a). As was the case for $M = 2$, aggregating discriminators evolve either in pure stable equilibria or in mixtures of two. Mixtures appear for both norms, but they are more common under scoring, where the potential for evolutionarily stable mixtures has previously been shown[9] (Supplementary Information Fig. 1). These results are qualitatively robust across a wide range of benefit-to-cost ratios and error rates (Extended Data Figs. 3 and 4), but the tolerance level that evolves depends on parameters: discriminators evolve to forgive more (lower $q$) when individuals are more error-prone or cooperation is less costly (higher benefit-to-cost ratio); conversely, they evolve to forgive less when assessments are more accurate or cooperation is costlier. The advantage of a certain level of tolerance $q$ can be understood by considering how different thresholds view the rest of the population (Fig. 3b). The more tolerant the discriminator, the more readily it assigns good reputations, which allows it to cooperate and be reciprocated with when error rates are higher. Conversely, the stricter the discriminator, the less likely it is to assign good reputations, which allows it to defect more and gain an advantage at low benefit-to-cost ratios. Moderately tolerant discriminators are more discerning, which is optimal in the remaining regimes (Extended Data Figs. 3 and 4).

Under both scoring and stern-judging, discriminators have a larger basin of attraction when individuals make more than two observations ($M > 2$), compared to when they make exactly two observations (Fig. 3a). These results suggest that increasing the number of observations might be better for cooperation. However, they do not guarantee that individuals will actually evolve to use more observations, given the choice. To investigate the evolution of $M$, we assume that information is freely available and that there are no cognitive constraints on using more observations. We fix $q$ and allow $M$ to vary, which results in a strategy space consisting of several discriminators $DISC_{q,M}$, along with ALLD and ALLC. When we allow $M$ to evolve by the same payoff-biased imitation process, we find that, regardless of the norm, the population evolves to use fewer observations than there are available (Fig. 3c). In fact, the population generally evolves to use the lowest number $M \geq 2$ of observations available; occasionally, there is coexistence of discriminators using the two lowest values of $M$ available. These results are qualitatively robust across a wide range of benefit-to-cost ratios and error rates (Extended Data Figs. 6 and 7), although larger values of $M$ may evolve for very high error rates or high benefit-to-cost ratios.

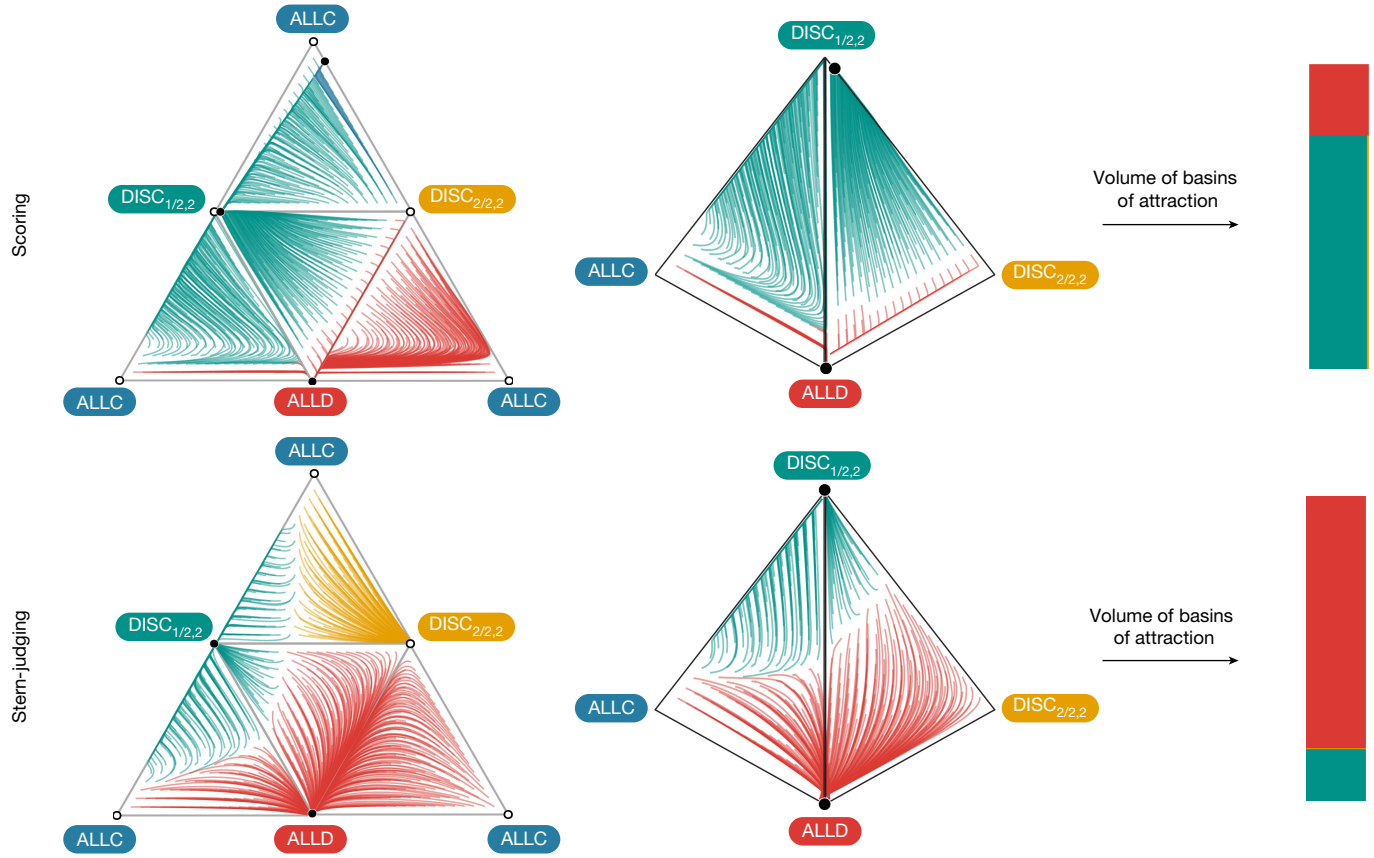



**Fig. 2 | For $M = 2$ observations, aggregating discriminators evolve even when considering the full strategy space.** We show the dynamics on the three-simplex (tetrahedron) whose four vertices correspond to the two unconditional strategies (ALLC and ALLD) and the two discriminating strategies ($DISC_{1/2,2}$ and $DISC_{2/2,2}$). The first panel on each row (corresponding to different norms) shows the dynamics on the faces of the tetrahedron. The volumes of the basins of attraction within the three-simplex (middle panel) are summarized in a bar, as shown by the arrow. The basins are estimated by numerical integration of 975 trajectories from evenly distributed initial frequencies in the interior of the three-simplex (Methods). The bars are formed by concatenating horizontal slices, each corresponding to the steady state reached from one of the 975 initial conditions. Horizontal slices with one colour correspond to pure equilibria; horizontal slices with two colours correspond to mixtures, with equilibrium frequencies reflected in the colour proportions. Note the difference between the dynamics on the faces (among triplets) versus the full strategy space. The rate of cooperation at the equilibria where the population consists entirely of discriminator strategies is 99.5% for scoring and 99.8% for stern-judging. Parameters: $b/c = 5$, $\alpha = 0.02$ and $\varepsilon = 0.02$. See corresponding results for simple-standing and shunning in Extended Data Fig. 2.

When we fix $q = 1/2$ and not-too-high error rates and benefit-to-cost ratios, the only stable discriminating equilibrium is pure 'look twice, forgive once' ($DISC_{1/2,2}$), regardless of the norm.

Our findings suggest that when individuals have the option to tune the number of observations they make, they do not generally evolve to use all available information. This outcome is counter-intuitive, especially because information is costless to harness in our model. Indeed, a more informed strategy can more accurately distinguish between justified and unjustified defections, so it assigns itself a higher reputation, leading to more self-cooperation (compare the three discs in Fig. 3d). Nevertheless, greater accuracy comes with an inherent cost: it makes a more informed strategy (higher $M$) weaker against strategies with a coarser view of the world (lower $M$). To see why, consider the competition between two such strategies with a fixed $q$: the more informed strategy is more likely to assign good reputations (Supplementary Information Fig. 2), which allows the less informed strategy to defect occasionally without incurring a reputational cost (see solid curves in Fig. 3d). Thus, there is a trade-off between the benefit of cooperating frequently with one's own strategy, which depends on the accuracy of judgements, and the cost of being too forgiving towards coarser discriminators. This trade-off divides the parameter space into two primary outcomes. For error rates and benefit-to-cost ratios that are not too high, the lowest possible $M \geq 2$ evolves (for example, $M = 2$ when $q = 1/2$); when either error rates or benefit-to-cost ratios are high, there are regions where consecutive values of $M$ coexist, separated by small regions in which only one $M$ dominates (Extended Data Fig. 8).

So far we have studied the evolution of the aggregation rule ($q$, $M$) by fixing one of its elements ($q$ or $M$) and varying the other. We now allow both $q$ and $M$ to evolve simultaneously. Due to computational complexity, we do so in a restricted strategy space: we pair $M = 2$ with one of $M = 4$, 6 or 8. We find that 'look twice, forgive once' is still selected, typically in a mixture with strategies that use more information and are stricter ($M > 2$ and $q > 1/2$). These strategies are, by themselves, unstable in competition against ALLC and ALLD. However, when 'look twice, forgive once' is present, their mixed equilibrium is stable against invasion by unconditional strategies (Extended Data Fig. 9). In these cases, some portion of the population will use the maximum amount of information available, while the rest will continue to use only a fraction of it. Although our theoretical result that individuals do not always evolve to use all freely available information may appear broadly counter-intuitive, it is supported by experimental literature[40] that emphasizes that humans waste information obtained by observing others or by communicating with them.

We have seen that tolerant judgements based on several observations produce robust agreement and cooperation, regardless of

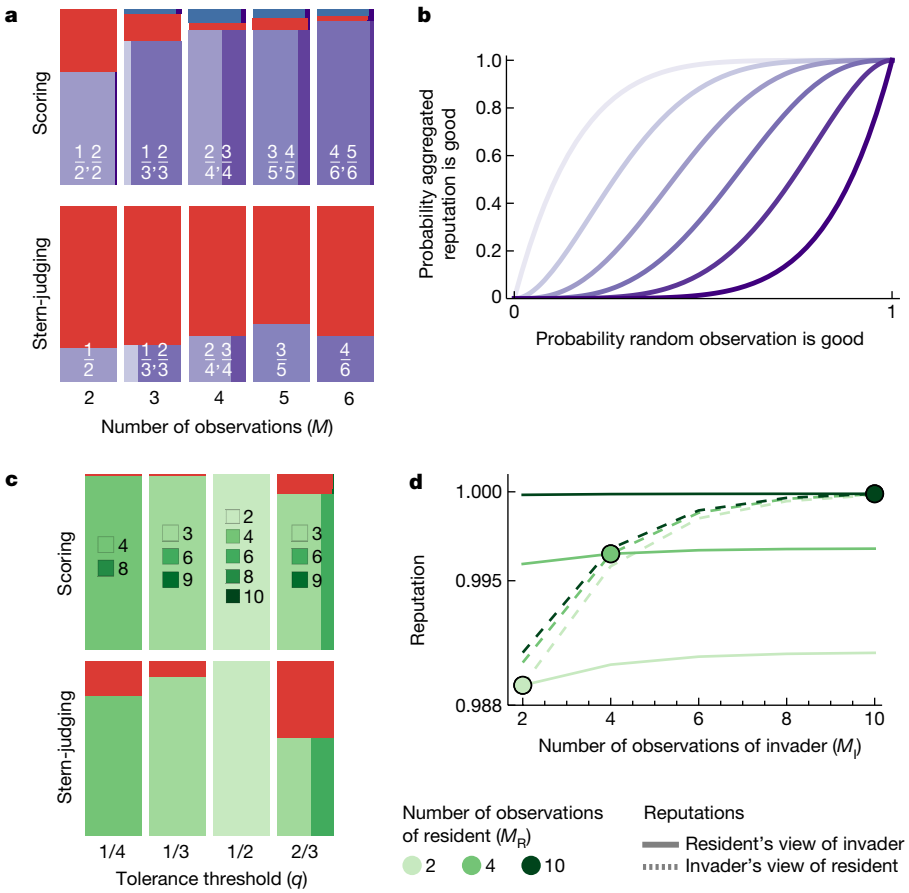



**Fig. 3 | Aggregating discriminators evolve for any number of observations $M \geq 2$, but, given the choice, individuals do not evolve to use all available information. a**, The evolution of tolerance ($q$), with all the possible aggregating discriminators for a fixed number of observations ($M$). Each bar summarizes the steady states reached by numerical integration from 100 different random initial conditions, for competition between unconditional strategies (ALLC and ALLD) and a set of aggregating discriminators. White labels indicate the aggregating discriminators (pure or mixed) that are stable against invasion by unconditional strategies. The rate of cooperation at the discriminating equilibrium ranges from 99.2% to 99.9% for both norms. **b**, The probability that an aggregating discriminator assigns a good reputation to someone after sampling $M$ observations (equation (6)) as a function of the probability that a single random observation is assessed as good (equation (4)). The panel shows this relationship for all strictness thresholds under the scoring norm with $M = 6$ observations. **c**, Evolution of the number of observations ($M$), with several aggregating discriminators using the same strictness threshold ($q$) but different numbers of observations ($M$). For each threshold, we analysed all integers $2 \leq M \leq 10$ that are multiples of $1/q$. Squares show the specific values of $M$ competing for each fixed $q$. The rate of cooperation at the equilibria where the population consists entirely of discriminator strategies ranges from 99.4% to 99.9% for both norms. **d**, The reputations that two discriminators with tolerance $q = 1/2$ but different values of $M$ (a resident using $M_R$ and an invader using $M_I$) assign to each other under the scoring norm. Continuous (dashed) lines represent the resident's (invader's) views of others. Discs indicate the resident's view of itself. For all panels, $b/c = 5$, $\alpha = 0.02$ and $\varepsilon = 0.02$.

the social norm a population adopts for evaluating behaviour. However, disagreements could still arise—especially in a multicultural society—from variation in norms[41]. To study the effects of norm heterogeneity, we analysed the co-evolution of strategies and norms. An individual's type is given by a strategy and a norm; during imitation, an individual can copy either the strategy or the norm of the role model. Interactions remain well mixed (Fig. 4a), and an individual's norm is unknown to others (individuals cannot condition their behaviour based on the recipient's norm). Competition occurs among the eight types resulting from pairing one of the four strategies (ALLC, ALLD, 'look twice, forgive once' or $DISC_{2/2,2}$) with one of the two norms (scoring or stern-judging; see Extended Data Fig. 10 for all other norm pairs). Regardless of how often individuals imitate norms versus strategies, we find that aggregating discriminators still have a substantial basin of attraction (Fig. 4b). When norms evolve, they cannot coexist indefinitely[42]; so, norm heterogeneity is transient. Notably, however, even when norm heterogeneity is exogenously fixed and, thus, the potential for norm-induced disagreements persists indefinitely, aggregating discriminators still have substantial basins of attraction. Intuitively, disagreement arising from norm variation is not fatal to aggregating discriminators because both norms agree that it is good to cooperate with a good person and bad to defect against a good person; so, there is some agreement between norms and, when at first they disagree, a second, forgiving look can facilitate reconciliation.

Overall, we have shown that both fundamental dilemmas of indirect reciprocity—the scoring and punishment dilemmas[27]—can be solved if individuals assign reputations by aggregating more than one observation, but not too many, while having some tolerance for bad actions. 'Look twice, forgive once' is the simplest example of such an aggregation rule. Such rules evolve for two reasons: on the one hand, they are tolerant enough to forgive mistakes and some intentional defections, which allows them to outperform classic discriminators and outcompete unconditional strategies; on the other hand, they are informationally coarse enough to defect a few times against themselves and against more informed strategies without incurring reputational cost. Not only do such strategies evolve, but they withstand challenges from norm heterogeneity.

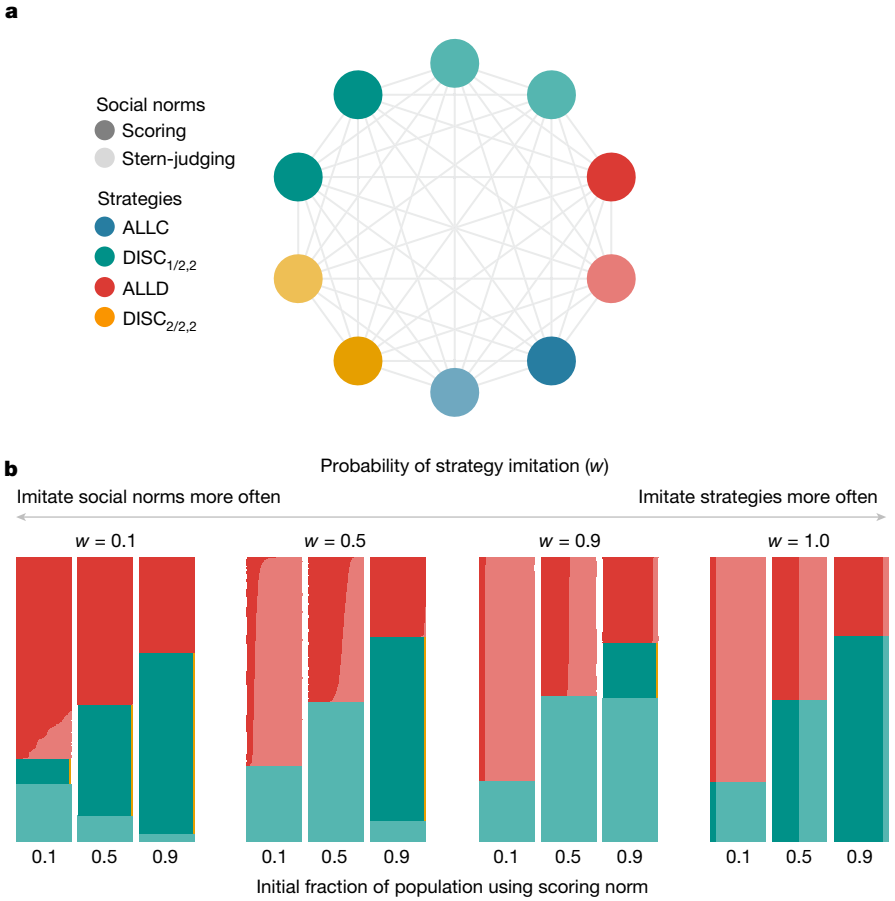



**Fig. 4 | Aggregating discriminators evolve even when there are several social norms of judgement. a**, Example of a population snapshot. Each individual (circle) has both a strategy (reflected by its hue) and a social norm (differentiated by its saturation; dark hues correspond to scoring and light hues correspond to stern-judging). Interaction and imitation are both well mixed. At each imitation step, an individual copies either the strategy (with probability $w$) or the social norm (with probability $1-w$) of a role model. **b**, The result of strategy and social norm co-evolution when an initial fraction of the population uses scoring and the remainder uses stern-judging (for example, 0.1 means that 10% of the population is initially assigned the scoring norm). Each set of three bars corresponds to a fixed probability of strategy imitation $w$ (for example, $w = 0.9$ means that 90% of the time individuals imitate strategies and the remaining 10% they imitate social norms; $w = 1$ means that individuals only copy strategies and so norm heterogeneity persists indefinitely). Each bar summarizes the steady states reached from 975 different initial conditions, by numerical integration of competition among eight types—four strategies (ALLC, ALLD, $\mathrm{DISC}_{1/2,2}$ or $\mathrm{DISC}_{2/2,2}$) paired with either scoring or stern-judging. The rate of cooperation at the equilibria where the population consists entirely of discriminator strategies exceeds 97.5% in all cases. Parameters: $b/c = 5$, $\alpha = 0.02$ and $\varepsilon = 0.02$.

These results have broader implications. Although our work was partly motivated by the dramatic increase in publicly available information, especially through social media, the mechanism we uncovered may be evolutionarily very old. This is because the mechanism works even under a simple, first-order social norm and because individuals ultimately evolve to use just a fraction of the available information, even when it is freely accessible. In other words, a strategy like 'look twice, forgive once' may have played a significant role in promoting cooperation even in small, early human groups that lacked complex rules for moral judgement or public institutions. As social groups became more sophisticated and complex social norms began to emerge--and even as groups became larger and harboured variation in social norms, --this simple heuristic could have continued to sustain cooperation under private information, without recourse to enforced agreement by a public institution. Altogether, these findings suggest that looking more than once and showing some tolerance may be fundamental components of social heuristics, which could have evolved early in human culture and persisted over time. This offers some hope for times when empathy between individuals and trust in institutions are eroded and norm heterogeneity is unavoidable: some cooperation and social function might, nonetheless, be maintained because, when all else fails, individuals may reflexively fall back on a simple and successful heuristic like 'look twice, forgive once'.

## Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-024-07977-x.


1. Roberts, G. et al. The benefits of being seen to help others: indirect reciprocity and reputation-based partner choice. *Philos. Trans. R. Soc. B* **376**, 20200290 (2021).
2. Raihani, N. *The Social Instinct: How Cooperation Shaped the World* (Random House, 2021).
3. Alexander, R. D. *The Biology of Moral Systems* (Routledge, 2017).
4. Nowak, M. A. & Sigmund, K. Evolution of indirect reciprocity. *Nature* **437**, 1291–1298 (2005).
5. Okada, I. A review of theoretical studies on indirect reciprocity. *Games* **11**, 27 (2020).
6. Ohtsuki, H. & Iwasa, Y. How should we define goodness?–reputation dynamics in indirect reciprocity. *J. Theor. Biol.* **231**, 107–120 (2004).
7. Hilbe, C., Schmid, L., Tkadlec, J., Chatterjee, K. & Nowak, M. A. Indirect reciprocity with private, noisy, and incomplete information. *Proc. Natl Acad. Sci. USA* **115**, 12241–12246 (2018).
8. Berger, U. Learning to cooperate via indirect reciprocity. *Games Econ. Behav.* **72**, 30–37 (2011).

9. Berger, U. & Grune, A. On the stability of cooperation under indirect reciprocity with first-order information. *Games Econ. Behav.* **98**, 19–33 (2016).
10. Schmid, L., Ekbatani, F., Hilbe, C. & Chatterjee, K. Quantitative assessment can stabilize indirect reciprocity under imperfect information. *Nat. Commun.* **14**, 2086 (2023).
11. Fujimoto, Y. & Ohtsuki, H. Evolutionary stability of cooperation in indirect reciprocity under noisy and private assessment. *Proc. Natl Acad. Sci. USA* **120**, e2300544120 (2023).
12. Radzvilavicius, A. L., Stewart, A. J. & Plotkin, J. B. Evolution of empathetic moral evaluation. *eLife* **8**, e44269 (2019).
13. Nakamaru, M. & Kawata, M. Evolution of rumours that discriminate lying defectors. *Evol. Ecol. Res.* **6**, 261–283 (2004).
14. Sommerfeld, R. D., Krambeck, H.-J., Semmann, D. & Milinski, M. Gossip as an alternative for direct observation in games of indirect reciprocity. *Proc. Natl Acad. Sci. USA* **104**, 17435–17440 (2007).
15. Seki, M. & Nakamaru, M. A model for gossip-mediated evolution of altruism with various types of false information by speakers and assessment by listeners. *J. Theor. Biol.* **407**, 90–105 (2016).
16. Radzvilavicius, A. L., Kessinger, T. A. & Plotkin, J. B. Adherence to public institutions that foster cooperation. *Nat. Commun.* **12**, 3567 (2021).
17. Iyengar, S., Lelkes, Y., Levendusky, M., Malhotra, N. & Westwood, S. J. The origins and consequences of affective polarization in the United States. *Annu. Rev. Political Sci.* **22**, 129–146 (2019).
18. Newton, K. & Norris, P. in *Disaffected Democracies: What's Troubling the Trilateral Countries* (eds Pharr, S. J. & Putnam, R. D.) 52–73 (Cambridge Univ. Press, 2000).
19. Public Trust in Government: 1958–2024. *Pew Research Center* www.pewresearch.org/politics/2024/06/24/public-trust-in-government-1958-2024/ (2024).
20. Boyd, R. & Richerson, P. J. The evolution of indirect reciprocity. *Soc. Netw.* **11**, 213–236 (1989).
21. Nowak, M. A. & Sigmund, K. The dynamics of indirect reciprocity. *J. Theor. Biol.* **194**, 561–574 (1998).
22. Manrique, H. M. et al. The psychological foundations of reputation-based cooperation. *Philos. Trans. R. Soc. B* **376**, 20200287 (2021).
23. Panchanathan, K. & Boyd, R. A tale of two defectors: the importance of standing for evolution of indirect reciprocity. *J. Theor. Biol.* **224**, 115–126 (2003).
24. Nowak, M. A. & Sigmund, K. Evolution of indirect reciprocity by image scoring. *Nature* **393**, 573–577 (1998).
25. Wedekind, C. & Milinski, M. Cooperation through image scoring in humans. *Science* **288**, 850–852 (2000).
26. Milinski, M., Semmann, D., Bakker, T. C. & Krambeck, H.-J. Cooperation through indirect reciprocity: image scoring or standing strategy? *Proc. R. Soc. B: Biol. Sci.* **268**, 2495–2501 (2001).
27. Okada, I. Two ways to overcome the three social dilemmas of indirect reciprocity. *Sci. Rep.* **10**, 16799 (2020).
28. Brandt, H. & Sigmund, K. The logic of reprobation: assessment and action rules for indirect reciprocation. *J. Theor. Biol.* **231**, 475–486 (2004).
29. 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).
30. Raihani, N. J. & Bshary, R. The reputation of punishers. *Trends Ecol. Evol.* **30**, 98–103 (2015).
31. Pacheco, J. M., Santos, F. C. & Chalub, F. A. C. Stern-judging: a simple, successful norm which promotes cooperation under indirect reciprocity. *PLoS Comput. Biol.* **2**, e178 (2006).
32. Ohtsuki, H. & Iwasa, Y. The leading eight: social norms that can maintain cooperation by indirect reciprocity. *J. Theor. Biol.* **239**, 435–444 (2006).
33. Takahashi, N. & Mashima, R. The importance of subjectivity in perceptual errors on the emergence of indirect reciprocity. *J. Theor. Biol.* **243**, 418–436 (2006).
34. Uchida, S. & Sasaki, T. Effect of assessment error and private information on stern-judging in indirect reciprocity. *Chaos Solitons Fractals* **56**, 175–180 (2013).
35. Uchida, S. Effect of private information on indirect reciprocity. *Phys. Rev. E* **82**, 036111 (2010).
36. Nowak, M. A. *Evolutionary Dynamics: Exploring the Equations of Life* (Harvard Univ. Press, 2006).
37. Sasaki, T., Okada, I. & Nakai, Y. The evolution of conditional moral assessment in indirect reciprocity. *Sci. Rep.* **7**, 41870 (2017).
38. Taylor, P. D. & Jonker, L. B. Evolutionary stable strategies and game dynamics. *Math. Biosci.* **40**, 145–156 (1978).
39. Boyd, R. & Richerson, P. J. *Culture and the Evolutionary Process* (Univ. of Chicago Press, 1985).
40. Morin, O., Jacquet, P. O., Vaesen, K. & Acerbi, A. Social information use and social information waste. *Philos. Trans. R. Soc. B* **376**, 20200052 (2021).
41. Henrich, J. & Gil-White, F. J. The evolution of prestige: freely conferred deference as a mechanism for enhancing the benefits of cultural transmission. *Evol. Hum. Behav.* **22**, 165–196 (2001).
42. Kessinger, T. A., Tarnita, C. E. & Plotkin, J. B. Evolution of norms for judging social behavior. *Proc. Natl Acad. Sci. USA* **120**, e2219480120 (2023).

# Methods

### Model description

**Interactions.** We consider a well-mixed population of individuals engaged in pairwise, one-shot interactions that take the form of donation games, also known as the simplified prisoner's dilemma. In each interaction, one individual is the potential donor of a cooperative act, and the other is the recipient. A donor can either cooperate (C), providing a benefit $b$ to the recipient at their own cost $c$ ($0 < c < b$), or defect (D), incurring no cost and providing no benefit to the recipient.

**Strategies.** The donor's action is determined by their view of the recipient's reputation. We model such reputations as binary values 0 or 1, which we refer to as bad (B) or good (G), respectively. We describe the strategies by the tuple $s = (s^B, s^G)$, where $s^R$ represents the decision to donate or not to a recipient with reputation $R \in \{B, G\}$. We consider three types of strategies: individuals who always cooperate (ALLC) with $s_{ALLC} = (1,1)$, individuals who always defect (ALLD) with $s_{ALLD} = (0,0)$ and individuals who discriminate by cooperating only with those in good standing ($DISC_{q,M}$) with $s_{DISC_{q,M}} = (0,1)$.

**Errors.** We allow errors in the execution of the strategy and in the assessment of each observed action[37]. With probability $\varepsilon$, an intended cooperative act is executed as defection, whereas an intended defection is always executed correctly (equation (3)). With probability $\alpha$, an observer incorrectly evaluates an action (equation (5)).

**Replicator dynamics.** We describe the dynamics of strategy competition in the limit of an infinitely large population using replicator equations[38]. Let $f_i$ represent the frequency of strategy $i$. The evolutionary dynamics under payoff-biased imitation follow the equations

$$\frac{df_i}{dt} = f_i\left(\Pi_i - \sum_{s\in S} f_s \Pi_s\right). \quad (1)$$

Here $S$ is the set of all strategies, and $\Pi_i$ is the average payoff of strategy $i$, computed as

$$\Pi_i = \sum_{j\in S} f_j\,(b a_{ji} - c a_{ij}), \quad (2)$$

where $ba_{ji}$ is the benefit of the donation made by $j$ and received by $i$, and $ca_{ij}$ is the cost of the donation made by $i$ and received by $j$. The term $a_{ij}$ is the expected action of a player with strategy $i$ towards a player with strategy $j$, given by

$$a_{ij} = (1-\epsilon)\,(s_i^B(1-r_{ij}^*) + s_i^G r_{ij}^*), \quad (3)$$

where $r_{ij}^*$ is the average reputation of strategy $j$ in the eyes of $i$ at the equilibrium of the reputation dynamics, as described below. Because $r_{ij}^*$ is an average, it may assume any value between 0 and 1, unlike the individual views of others' reputations, which are binary.

**Reputation updates.** After everyone has played the game with everyone else, individuals update their views of every population member. Reputations are tracked according to a solitary private monitoring system, whereby all individuals observe and update their views of others independently[7,43]. To assign reputations, individuals observe the actions of others and assess each action as good or bad according to a social norm.

**Social norms.** We describe a social norm by a vector $\mathbf{d} = \{d_{CG}, d_{CB}, d_{DG}, d_{DB}\} \in \{0,1\}$ (ref. 4), where the entry $d_{AR}$ denotes whether an action $A \in \{C, D\}$ towards someone with reputation $R \in \{B, G\}$ is considered bad or good. For example, $d_{CG} = 1$ means that cooperating with good recipients is perceived as good, and $d_{CB} = 0$ means that cooperating with bad recipients is considered bad. Previous studies have shown that social norms that foster cooperation and are evolutionarily stable share the principle of rewarding cooperation and punishing defection towards good individuals[6]. We explore the social norms that follow this principle but vary in their assessment of actions towards individuals with a bad reputation: scoring[24,25] ($d = \{1,1,0,0\}$), stern-judging[31,44,45] ($d = \{1,0,0,1\}$), simple-standing[26,46] ($d = \{1,1,0,1\}$) and shunning[33] ($d = \{1,0,0,0\}$).

**Observations.** To assign a binary reputation based on several observed actions, each of which is assessed as good or bad according to the observer's social norm, the observer must somehow aggregate information across observations. We study an aggregation rule defined by a number of observations $M$ and a strictness threshold $q \in [0,1]$, which determines the minimum number of good actions ($\lceil qM \rceil$ out of the $M$ observed) needed to confer a good reputation.

The probability that an observer with strategy $i$ assesses as good a single randomly selected interaction between a donor with strategy $j$ and a recipient with strategy $k$ is

$$p_{ijk} = a_{jk}\,(r_{ik} d_{CG} + (1-r_{ik})\,d_{CB}) + (1-a_{jk})\,(r_{ik} d_{DG} + (1-r_{ik})\,d_{DB}), \quad (4)$$

where $d_{AR}$ is prescribed by the social norm, $a_{jk}$ is the action of the donor and $r_{ik}$ is the reputation of the recipient in the eyes of the observer.

Then, the probability that an observer with strategy $i$ assesses as good an action of a donor with strategy $j$ towards a randomly selected recipient is

$$p_{ij} = \alpha + (1-2\alpha) \sum_{k\in S} f_k\, p_{ijk}. \quad (5)$$

An observer assesses $M$ actions of a donor and assigns the donor a good reputation if at least $qM$ assessments are good. We let $(q_i, M_i)$ be the aggregation rule corresponding to strategy $i$; in other words, a strategy-$i$ observer makes $M_i$ observations and uses a tolerance threshold $q_i$. Given that the $M_i$ observed interactions are chosen uniformly at random and with replacement, the probability $\rho_{ij}$ that a strategy-$i$ observer assigns a strategy-$j$ donor a good reputation—which is equivalent to the probability that the observer assesses at least $q_i M_i$ actions of a strategy-$j$ donor as good—is given by

$$\rho_{ij} = \sum_{m=\lceil q_i M_i \rceil}^{M_i} \binom{M_i}{m} p_{ij}^m (1-p_{ij})^{M_i - m}. \quad (6)$$

As is common in the literature on indirect reciprocity, we assume that reputation updates occur at a faster timescale than strategy updates, so that reputations equilibrate before individuals update their strategies[37,43,47]. In an infinitely large population, the change in average reputations $r_{ij}$ is very small after a single interaction, so we can approximate their dynamics using the ordinary differential equation[47] (see 'Reputation updating' in the Supplementary Information for a detailed derivation)

$$\frac{dr_{ij}}{dt} = \rho_{ij} - r_{ij}. \quad (7)$$

Note from equations (4–6) that $\rho_{ij}$ is a function of $r_{ij}$. The equilibrium values $r_{ij}^*$ such that $\left.\frac{dr_{ij}}{dt}\right|_{r_{ij}=r_{ij}^*} = 0$ are used in equation (3) and to integrate the replicator equations (equation (1)) numerically.

Previous work has assumed a unique average reputation equilibrium, regardless of the initial condition $r(t=0)$ (the average reputation at the onset of the strategic updates). Here, we find that this is not always the case. Instead, in certain parameter regimes, bistability is possible: different average reputation equilibria will be reached depending on whether $r(0)$ is closer to 0 or to 1 (whether individuals start with an initially pessimistic versus initially optimistic view of the world). For all our results, we assume that individuals start with an optimistic view.

Before the very first update of reputations, they start by regarding everyone as good $r(0) = 1$, which is in keeping with direct reciprocity assumptions that conditional strategies start by being 'nice'[48]. However, the bistability arises only for $M \geq 3$, and thus, the optimism assumption does not affect the robustness of 'look twice, forgive once'.

### Numerical integration

To numerically investigate the evolutionary dynamics of any set of strategies, we applied the following procedure:

1. Set initial frequencies for the strategies ('Sampling initial frequencies' in the Supplementary Information); set $r(0) = 1$ as discussed above.
2. Solve the reputation dynamics at equilibrium (equation (7)).
3. Integrate a time step of the replicator dynamics (equation (1)).
4. Repeat steps 2 and 3 until an equilibrium is reached.

We stopped integration when the absolute value of the derivative reached was less than or equal to $10^{-10}$, at which point we consider the system to have reached an equilibrium. To estimate the volume of the basin of attraction towards each strategy, we concatenated the steady states reached by the set of initial frequencies. To calculate the rate of cooperation at each discriminating steady state (equilibria without unconditional strategies), we solved for the equilibrium reputations $r_{ij}^*$ at the equilibrium frequencies $f_i^*$ and computed the weighted sum $\sum_i f_i^* \sum_j r_{ij}^* f_j^*$.

**Evolution of reputation use in the presence of unconditional strategies.** To study the dynamics of competition between discriminating and unconditional strategies, we uniformly sampled initial frequencies with the three strategies presented (ALLC, ALLD and $\mathrm{DISC}_{q,M}$) for different strictness thresholds $q$ and numbers of observations $M$. Specifically, for $M = 2$, we report the temporal trajectories starting from points in the interior of the two-simplex (triangle), where each strategy frequency $f_i \in \{0.05, 0.10, \ldots, 0.85, 0.90\}$. We show the trajectories on the two-simplexes in Extended Data Fig. 1 and on the two-dimensional-projected faces of the three-simplex (tetrahedron) in the first column of Fig. 2. Each triplet of strategies was run independently.

**Robustness of aggregating discriminators against an unconditional probabilistic strategy.** We explored the robustness of aggregating discriminators against an unconditional strategy that cooperates with probability $p$ and defects with probability $1 - p$, denoted $\mathrm{probC}_p$. To study this, we first considered pairwise competition of $\mathrm{probC}_p$ against homogeneous populations of discriminators (DISC, $\mathrm{DISC}_{1/2,2}$ and $\mathrm{DISC}_{2/2,2}$) for $p \in \{0.0, 0.1, \ldots, 0.9, 1.0\}$. We sampled 19 uniformly distributed initial conditions from the one-simplex (line) and integrated until the dynamics reached a steady state (Extended Data Fig. 3a–c). Next, we allowed for a larger strategy space for $M = 1$ that comprises $\mathrm{probC}_p$, ALLD and DISC and for $M = 2$ that comprises $\mathrm{probC}_p$, ALLD and the two aggregating discriminators ($\mathrm{DISC}_{1/2,2}$ and $\mathrm{DISC}_{2/2,2}$) for $p \in \{0.0, 0.1, \ldots, 0.9, 1.0\}$. We studied competition among such strategies by sampling 171 uniformly distributed initial frequencies from the interior of the two-simplex (for $M = 1$) and 1,000 random initial frequencies from the three-simplex (for $M = 2$) (Extended Data Fig. 3d,e).

**Evolution of the strictness threshold $q$.** To investigate the evolution of tolerance, we fixed the number of observations $M$ and allowed $q$ to evolve (by payoff-biased imitation). The strategy space included the two unconditional strategies (ALLC and ALLD) and all possible types of $\mathrm{DISC}_{q,M}$ with thresholds $q \in \{1/M, \ldots, M/M\}$ for a fixed $M$. For $M = 2$, we estimated the basin volumes by sampling 975 uniformly distributed initial frequencies from the interior of the three-simplex (tetrahedron) and followed the procedure described above (Fig. 2 right). For larger numbers of observations $M > 2$, we estimated the basin volumes by sampling 100 random initial frequencies with all strategies present (Fig. 3a and Extended Data Figs. 4 and 5).

**Evolution of the number of observations $M$.** To investigate the evolution of the number of observations, we fixed the strictness threshold $q$ and allowed $M$ to evolve (by payoff-biased imitation). The strategy space included the two unconditional strategies (ALLC and ALLD) and all the aggregating discriminators using numbers of observations $2 \leq M \leq 10$ that are integer multiples of $1/q$. For example, for $q = 1/3$, we included three aggregating discriminators using $M \in \{3, 6, 9\}$, whereas for $q = 1/2$, we included five aggregating discriminators using $M \in \{2, 4, 6, 8, 10\}$. We estimated basin volumes by sampling 100 random initial frequencies from the interior of the simplexes (Fig. 3c and Extended Data Figs. 6 and 7).

**Co-evolution of the number of observations $M$ and the strictness threshold $q$.** To investigate the evolution of $M$ and $q$ simultaneously (Extended Data Fig. 9), we fixed a social norm and selected a set of discriminating strategies that used different values of $M$ and $q$. We first identified the thresholds that evolve for a fixed $M \in \{4, 6, 8\}$ by running a competition among the possible thresholds for a fixed $M$. For example, we simulated the dynamics between the unconditional strategies and all the possible thresholds for $M = 4$, and we found that a mix between $q = 2/4$ and $q = 3/4$ has a basin of attraction (Extended Data Fig. 9a). Then, we included the thresholds that could coexist at equilibrium into a strategy space comprising two unconditional strategies (ALLC and ALLD) and the aggregating discriminators that evolve for $M = 2$. Following our example, we simulated the dynamics between the unconditional strategies (ALLC and ALLD) and four $\mathrm{DISC}_{q,M}$ strategies: two using $M = 2$ ($q = 1/2$ and $q = 2/2$) and two using $M = 4$ ($q = 2/4$ and $q = 3/4$) (Extended Data Fig. 9c). We repeated this procedure for $M = 6$ and $M = 8$ (Extended Data Fig. 9).

**Co-evolution of strategies and social norms.** To study the co-evolution of strategies and social norms (Fig. 4 and Extended Data Fig. 10), we allowed individuals to imitate strategies and social norms independently. At each imitation step, an individual might copy either the strategy of the role model, with probability $w$, or the norm of the role model, with probability $1 - w$ ('Norm dynamics' section in the Supplementary Information). For both Fig. 4 and Extended Data Fig. 10, we simulated competition among the eight types obtained by pairing one of the four strategies (ALLC, ALLD, $\mathrm{DISC}_{1/2,2}$ or $\mathrm{DISC}_{2/2,2}$) with one of two norms (scoring or stern-judging). We estimated the volumes of the basins of attraction by sampling 975 initial strategic frequencies and following the procedure described above. We considered four different values for the probability of strategy imitation $w \in \{0.1, 0.5, 0.9, 1.0\}$. For Fig. 4, we considered three different values for the initial fraction of the population using scoring (10%, 50% and 90%), with the remaining fraction using stern-judging. Extended Data Fig. 10 shows analogous results for every pair of norms, with each norm initially used by 50% of the population.

### Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

## Data availability

There are no empirical data associated with this study.

## Code availability

The code for generating the numerical calculations in this study is freely available at Zenodo (https://doi.org/10.5281/zenodo.12795781)[49] and through GitHub at github.com/michel-mata/IRMO.jl.

43. Okada, I., Sasaki, T. & Nakai, Y. A solution for private assessment in indirect reciprocity using solitary observation. *J. Theor. Biol.* **455**, 7–15 (2018).
44. Kandori, M. Social norms and community enforcement. *Rev. Econ. Stud.* **59**, 63–80 (1992).
45. Santos, F. P., Pacheco, J. M. & Santos, F. C. Evolution of cooperation under indirect reciprocity and arbitrary exploration rates. *Sci. Rep.* **6**, 37517 (2016).
46. Sugden, R. et al. *The Economics of Rights, Co-operation and Welfare* (Springer, 2004).
47. Perret, C., Krellner, M. & Han, T. A. The evolution of moral rules in a model of indirect reciprocity with private assessment. *Sci. Rep.* **11**, 23581 (2021).
48. Axelrod, R. & Hamilton, W. D. The evolution of cooperation. *Science* **211**, 1390–1396 (1981).
49. Michel-Mata, S. The evolution of private reputations in information-abundant landscapes. *Zenodo* https://doi.org/10.5281/zenodo.12795781 (2024).



**Acknowledgements** We thank members of the Tarnita and Plotkin labs and D. Ocampo for productive discussions. We thank J. Chisausky for testing the code for reproducibility. We are grateful to the anonymous reviewers for much constructive feedback. We acknowledge support from the James S. McDonnell Foundation, a Postdoctoral Fellowship Award in Understanding Dynamic and Multi-scale Systems (https://doi.org/10.37717/2021-3209 to M.K.), the Simons Foundation Math+X Grant to the University of Pennsylvania (J.B.P.) and the John Templeton Foundation (Grant No. 62281 to J.B.P. and T.A.K.).


**Author contributions** S.M.M., M.K., J.S., T.A.K., J.B.P. and C.E.T. conceived the study and developed and analysed the mathematical model. S.M.M., J.B.P. and C.E.T. wrote the paper with input from M.K., J.S. and T.A.K.

**Competing interests** The authors declare no competing interests.

**Additional information**

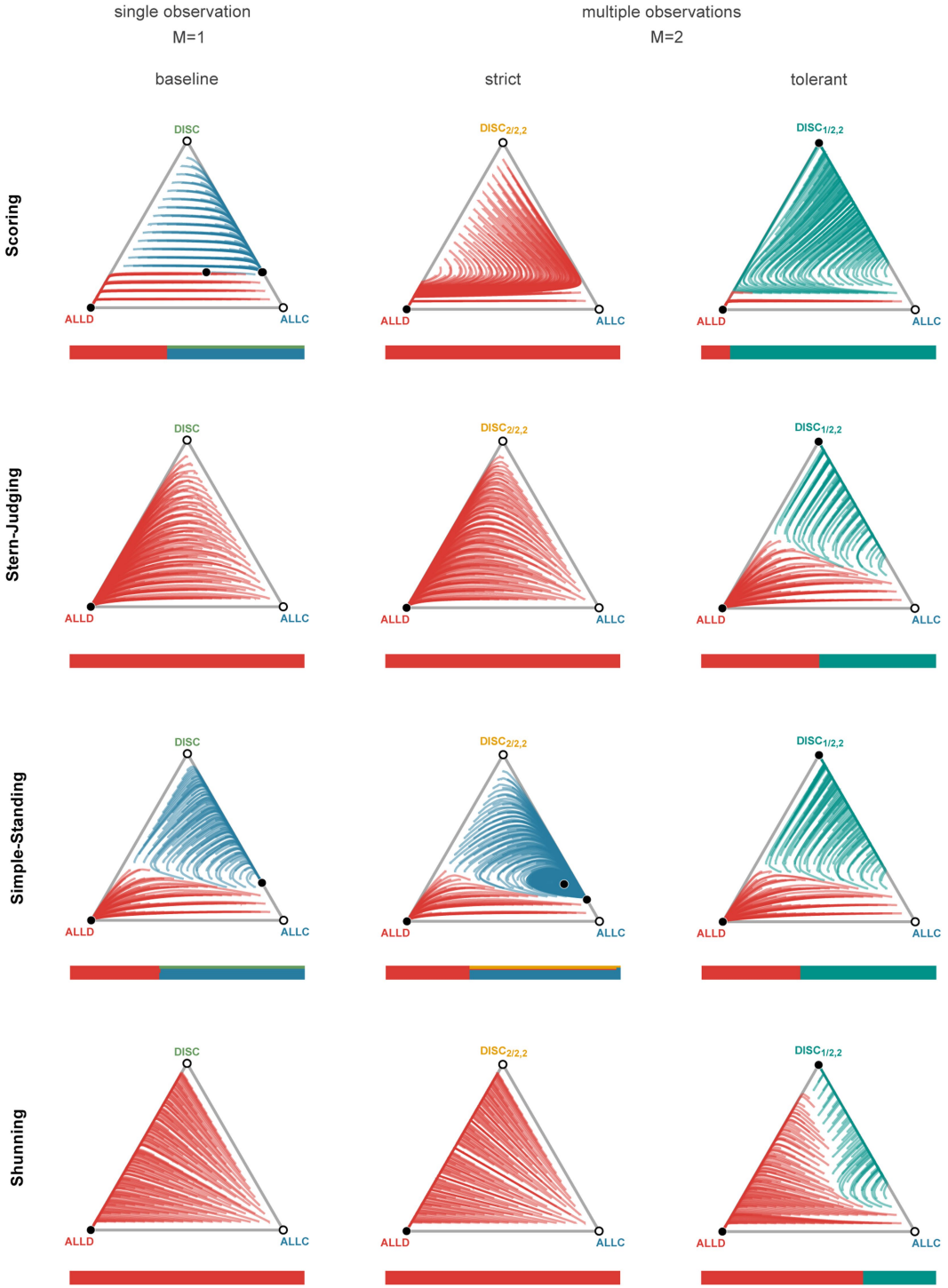

**Supplementary information** The online version contains supplementary material available at https://doi.org/10.1038/s41586-024-07977-x.



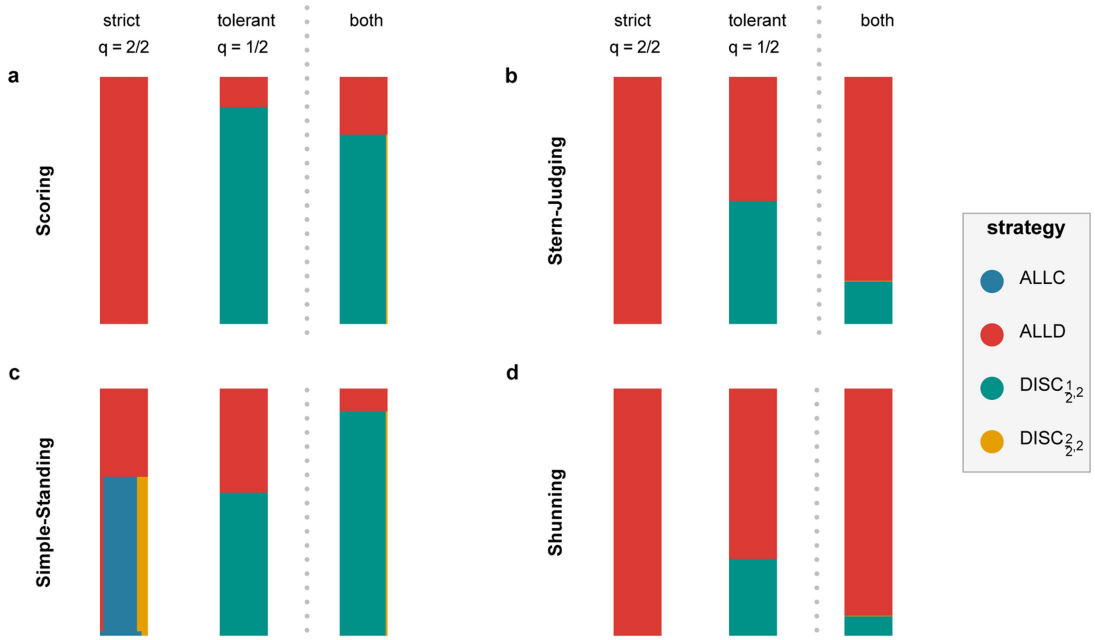
**Correspondence and requests for materials** should be addressed to Corina E. Tarnita.



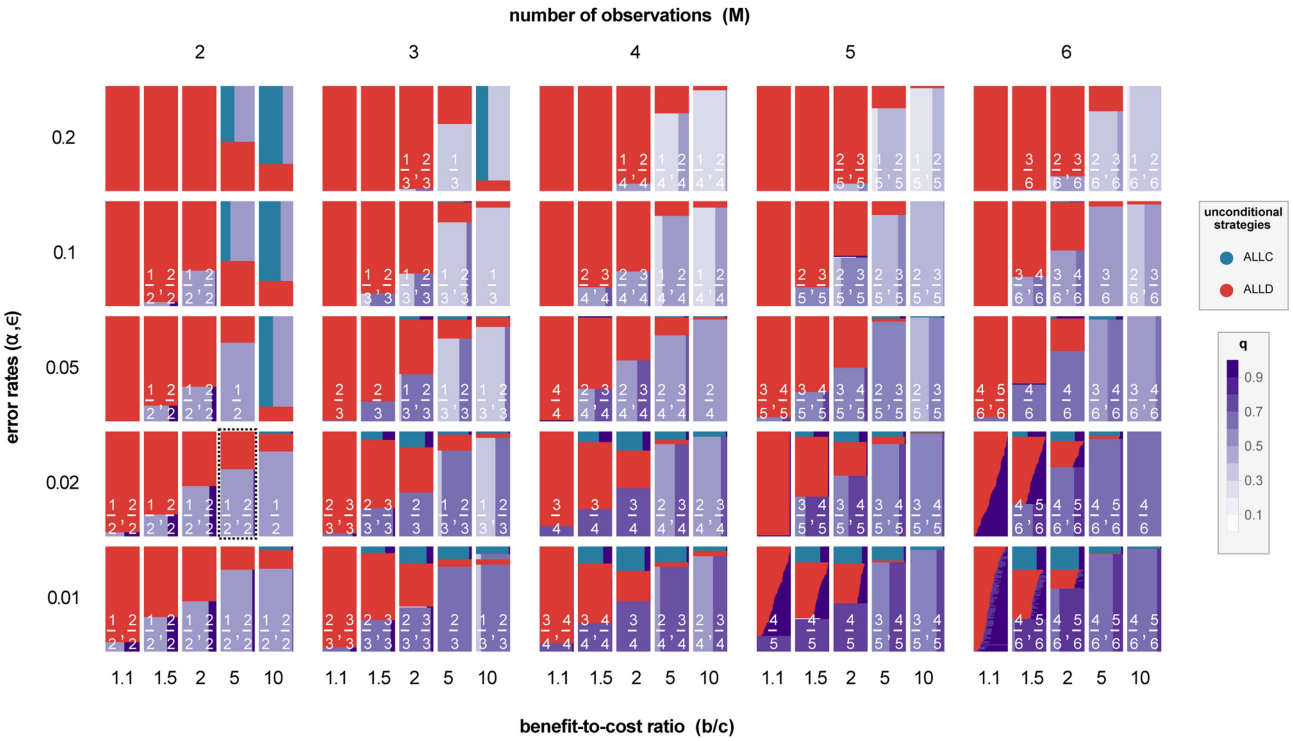
**Peer review information** *Nature* thanks the anonymous reviewers for their contribution to the peer review of this work.



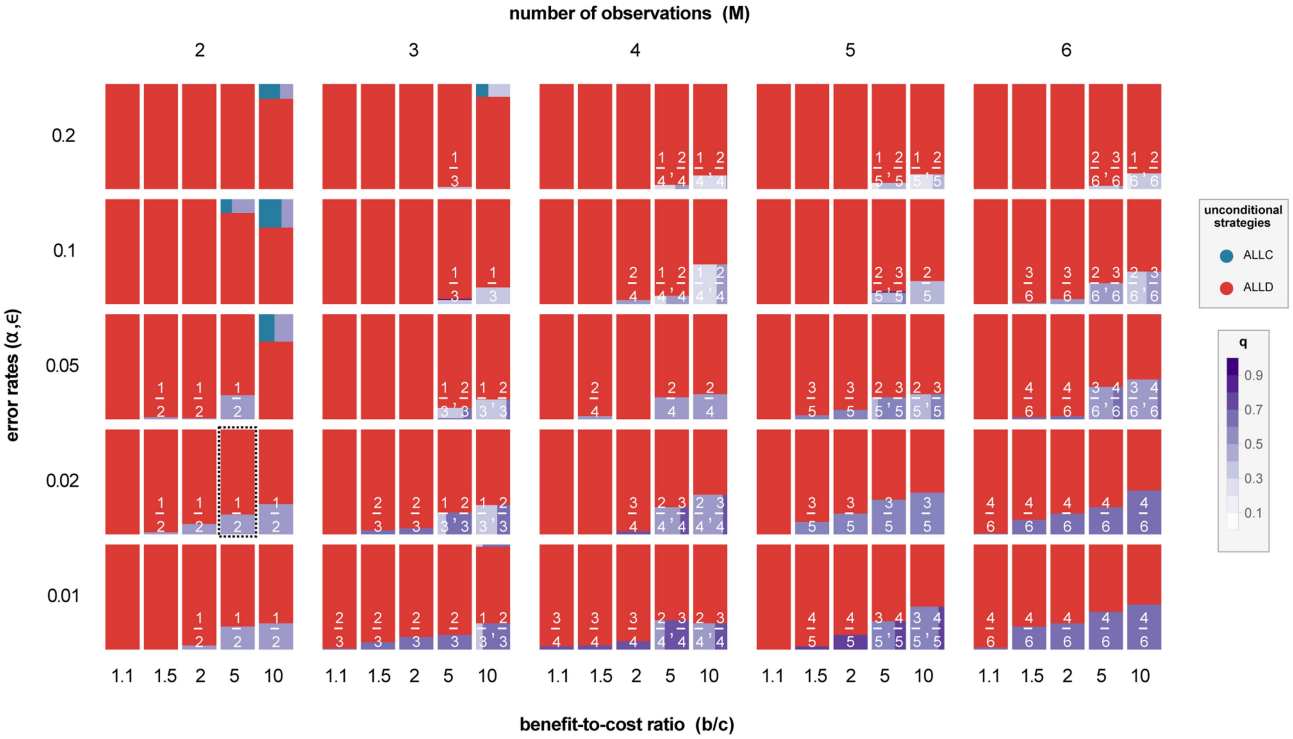

**Extended Data Fig. 1 | 'Look twice, forgive once' is stable and facilitates cooperative outcomes regardless of the social norm.** We show the dynamics between *ALLC*, *ALLD*, and a discriminator strategy using $M = 1$ or $M = 2$. The classical *DISC* ($M = 1$) is unstable for all norms, and the cooperative outcomes for scoring and simple-standing are vulnerable to invasion by *ALLD*. When $DISC_{q,M}$ uses $M = 2$ observations and is strict ($q = 2/2$), the outcomes of the dynamics are worse than with a single observation. But when $DISC_{q,M}$ is tolerant ($q = 1/2$), it is stable regardless of the social norm. The rate of cooperation at the 'look twice, forgive once' equilibrium is > 99.8% for all norms. For all panels, the benefit-to-cost ratio is $b/c = 5$, with error rates of assessment and execution $\alpha = 0.02$ and $\varepsilon = 0.02$, respectively.



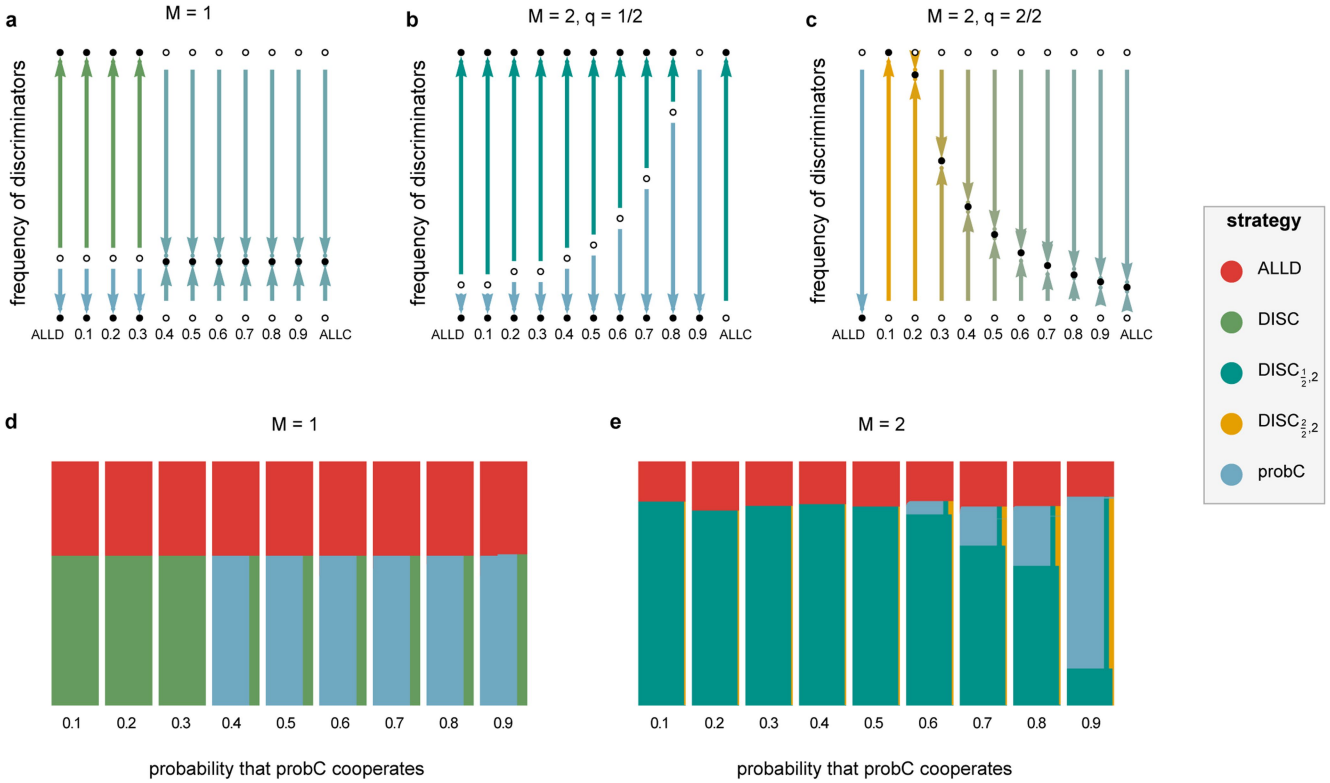

**Extended Data Fig. 2 | For $M = 2$, aggregating discriminators evolve versus unconditional strategies.** We estimated the volume of the basins of attraction towards aggregating discriminators, for a strategy space containing either strict ($DISC_{2/2,2}$) or tolerant ($DISC_{1/2,2}$) aggregating discriminators in the presence of *ALLC* and *ALLD* (left of dashed lines); and for the full strategy space with unconditional strategies and both aggregating discriminators (right of dashed lines). Each panel corresponds to a fixed, shared social norm. We estimated the basins by numerically integrating trajectories from evenly distributed initial frequencies in the interior of the simplex (171 for triplets and 975 for quartets; see Methods). The bars concatenate the steady states of all these trajectories. Regardless of the norm, tolerant discriminators are stable against unconditional strategies, while strict discriminators are not. When the full strategy space is considered, aggregating discriminators always have a basin of attraction. Under scoring or simple-standing, the stable equilibrium is a mix, with a large fraction of 'look twice, forgive once' coexisting with a small fraction of strict discriminators. Under stern-judging or shunning, we find a pure 'look twice, forgive once' stable equilibrium. The rate of cooperation at the discriminating equilibrium is 99.5% for scoring, 99.8% for stern-judging, 99.4% for simple-standing, and 99.8% for shunning. For all panels, the benefit-to-cost ratio is $b/c = 5$, with error rates of assessment and execution $\alpha = 0.02$ and $\varepsilon = 0.02$, respectively.



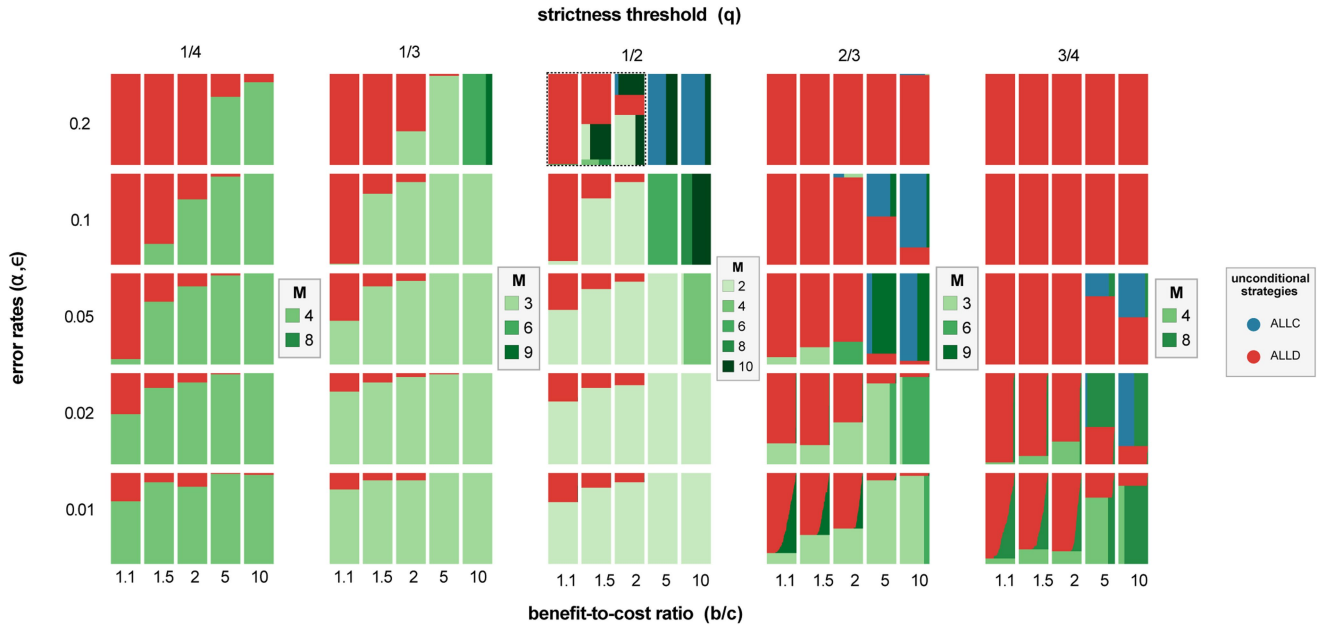

**Extended Data Fig. 3 | Robustness of the evolution of tolerance thresholds under scoring.** Panels show the competition of aggregating discriminators using all possible thresholds $q$ for a fixed number of observations $M$ in the presence of the unconditional strategies, for varying values of the benefit-to-cost ratio ($b/c$) and of the assessment ($\alpha$) and execution ($\varepsilon$) error rates (with $\alpha = \varepsilon$). Each bar concatenates the steady states reached by numerical integration from 100 different initial conditions. White labels indicate the aggregating discriminators (pure or mixed) that are stable against invasion by unconditional strategies. The rates of cooperation at the discriminating equilibria range from 69.2% to 99.9%. Of the 99 steady states shown in this plot, only 11 are below a rate of cooperation of 90%.



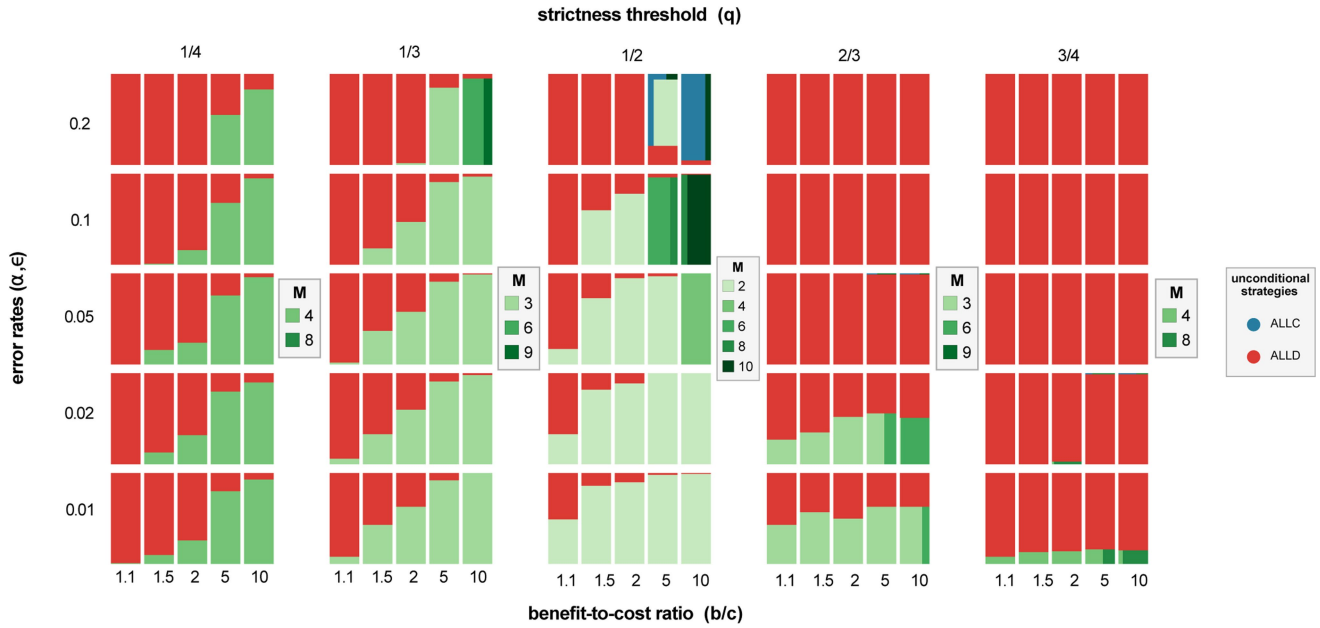

**Extended Data Fig. 4 | Robustness of the evolution of tolerance thresholds under stern-judging.** Panels show the competition of aggregating discriminators using all possible thresholds $q$ for a fixed number of observations $M$ in the presence of the unconditional strategies, for varying values of the benefit-to-cost ratio ($b/c$) and of the assessment ($\alpha$) and execution ($\varepsilon$) error rates (with $\alpha = \varepsilon$). Each bar concatenates the steady states reached by numerical integration from 100 different initial conditions. White labels indicate the aggregating discriminators (pure or mixed) that are stable against invasion by unconditional strategies. The rates of cooperation at the discriminating equilibria range from 30.1% to 99.9%. Of the 76 steady states shown in this plot, only 5 are below a rate of cooperation of 90%.



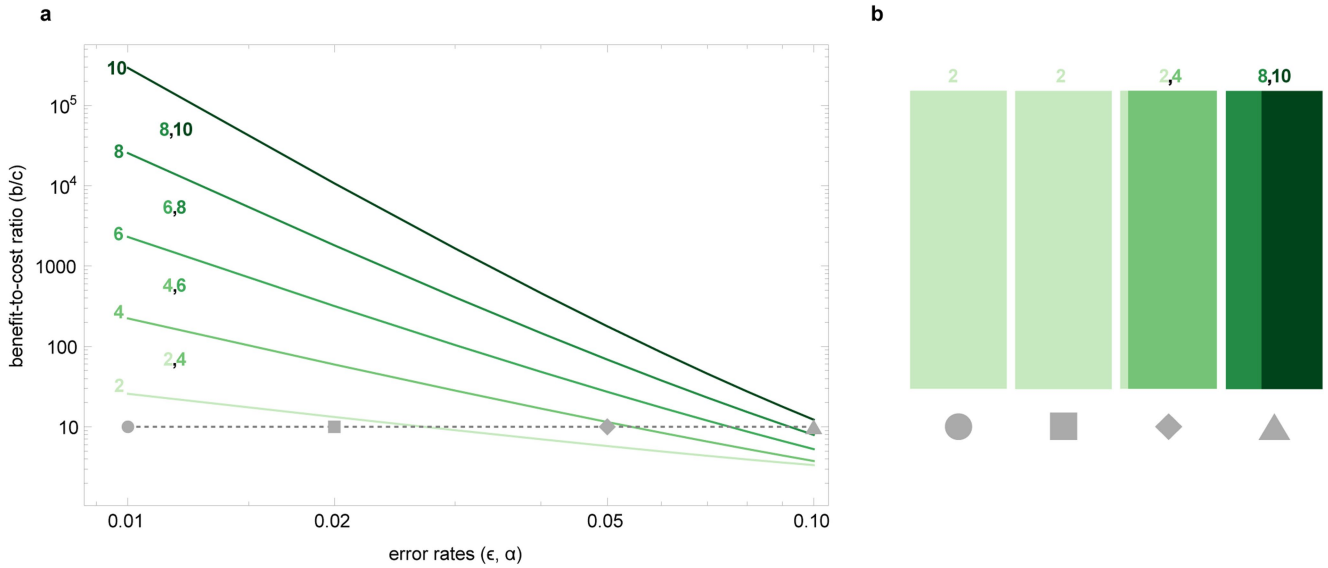

**Extended Data Fig. 5 | An unconditional strategy *probC*$_p$ can invade classical *DISC* and pure *DISC*$_{q,M}$, but not mixed equilibria. a-c**, Competition between probabilistic unconditional cooperator (*probC*$_p$), classic discriminators (*DISC*), and aggregating discriminators (*DISC*$_{q,M}$) under scoring. The horizontal axis shows multiple values of *probC*$_p$'s probability of cooperation ($p = 0$ being *ALLD* and $p = 1$ being *ALLC*). The vertical axes represent the frequency of discriminators. Arrows show the flow of the evolutionary dynamics along the vertical axis. Filled-in circles represent stable equilibria; open circles represent unstable equilibria. **a**, *probC*$_p$ can invade and coexist with the classic single-observation discriminator (*DISC*) when the probability of cooperating is large enough ($p > 0.3$). At low values of $p \leq 0.3$, coexistence becomes bistability. **b**, When competing against 'look twice, forgive once', only a narrow range of high probabilities ($0.8 < p < 1.0$) allows *probC*$_p$ to invade. At lower $p \leq 0.8$, we find bistability. **c**, *probC*$_p$ is also able to invade strict discriminators and coexist with them for $p > 0.1$, with a larger fraction of *probC*$_p$ at this mixed equilibrium as $p$ increases. **d-e**, Effects of *probC*$_p$ when more strategies are present. The bars show the results of the competition between *probC*$_p$, *ALLD*, and *DISC* (for $M = 1$); and the competition between *probC*$_p$, *ALLD*, tolerant aggregating discriminator ($q = 1/2$), and strict aggregating discriminator ($q = 2/2$) (for $M = 2$). **d**, When discriminators use a single observation, the outcomes are very similar to the classic scenario with *ALLC*: if $p$ is large enough ($p > 0.3$), there is coexistence between *probC*$_p$ and *DISC* with the same weakness of the 'scoring dilemma' (i.e. *ALLD* can invade and take over). **e**, With multiple observations, for all values of $p$ we tested, a mixture of tolerant and strict discriminators can coexist, have a non-trivial basin of attraction, and resist invasion by *probC*$_p$ and *ALLD*. This is because having a few strict discriminators in the mix effectively increases the overall strictness of the population compared to a scenario where only *DISC*$_{1/2,2}$ is present. As a result, the overall 'effective population tolerance' becomes $q_e > 1/2$. This higher level of strictness enables the population to more successfully identify and punish *probC*$_p$ when rare. As $p$ increases, the basin of attraction towards such mixed discriminating equilibrium decreases. For all panels, the benefit-to-cost ratio is $b/c = 5$, with error rates of assessment and execution $\alpha = 0.02$ and $\varepsilon = 0.02$, respectively.



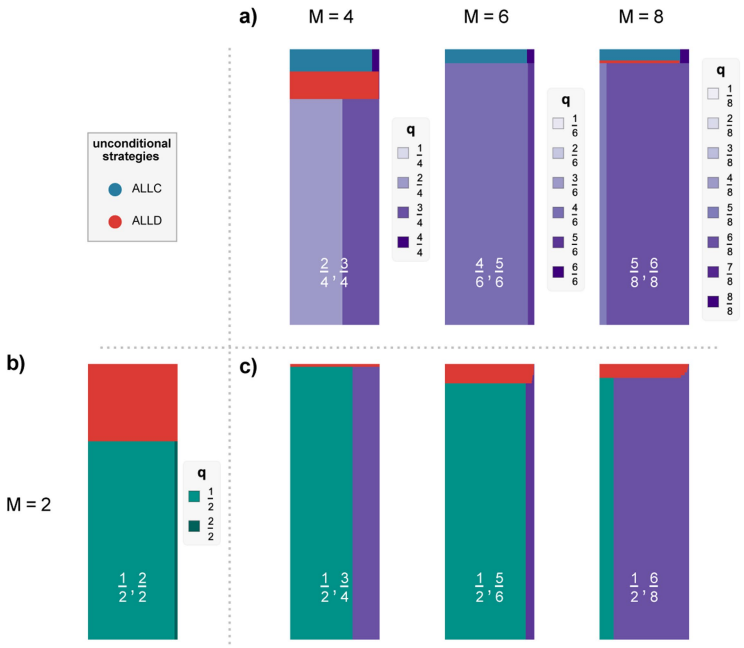

**Extended Data Fig. 6 | Robustness of the evolution of the number of observations under scoring.** Panels show the competition of all possible numbers of observations ($2 \le M \le 10$) for a fixed strictness threshold $q$ in the presence of unconditional strategies, for varying values of the benefit-to-cost ratio ($b/c$) and of the assessment ($\alpha$) and execution ($\varepsilon$) error rates (with $\alpha = \varepsilon$). Each bar concatenates the steady states reached by numerical integration from 100 different initial conditions. The rates of cooperation at the discriminating equilibria range from 79.2% to 100%. Of the 88 steady states shown in this plot, only 3 are below a rate of cooperation of 90% (marked with a dashed line).




**Extended Data Fig. 7 | Robustness of the evolution of the number of observations under stern-judging.** Panels show the competition of all possible numbers of observations ($2 \le M \le 10$) for a fixed strictness threshold $q$ in the presence of unconditional strategies, for varying values of the benefit-to-cost ratio ($b/c$) and of the assessment ($\alpha$) and execution ($\varepsilon$) error rates (with $\alpha = \varepsilon$). Each bar concatenates the steady states reached by numerical integration from 100 different initial conditions. The rates of cooperation at all 76 discriminating equilibria in this plot range from 92.4% to 100%.




**Extended Data Fig. 8 | Conditions for the evolution of a number of observations *M* under a fixed strictness threshold *q*.** **a**, The panel illustrates the criteria determining how a resident discriminator, which aggregates $M_R$ observations, can resist invasion by a discriminator that aggregates a larger number of observations ($M_I > M_R$). The criterion for each resident type is shown by distinct curves. Below each curve are the conditions where the resident discriminator successfully resists invasion, while above the curves indicates vulnerability to invasion by discriminators aggregating a larger number of observation. The space between adjacent curves is a coexistence zone, where both resident and invading discriminators can stably exist together. Lower numbers of observations are favored by lower benefit-to-cost ratios. As error rates increase, the benefit-to-cost ratio required by a resident to resist invasion decreases exponentially. The dashed line shows $b/c = 10$ as an example, with different markers indicating specific values of the error rates. **b**, Panel shows the outcomes of the competition between multiple aggregating discriminators using $M \in \{2,4,6,8,10\}$ in the presence of the unconditional strategies. Each bar correspond to each of the markers of panel **a**, showing the steady states reached by numerical integration from 100 different initial conditions. The steady states confirm the criteria of panel **a** and show that, as error rates increase with a fixed benefit-to-cost ratio, larger numbers of observations evolve. For all panels, we set $q = 1/2$ and fixed scoring as the social norm.




**Extended Data Fig. 9 | Coevolution of the number of observations *M* and the strictness threshold *q* under scoring. a**, Estimated basins of attraction after competition of all possible thresholds for $M \in \{4,6,8\}$ and the two unconditional strategies. The mix between two moderately tolerant discriminators has the largest basin of attraction regardless of *M*. **b**, Estimated basins of attraction after competition of all possible thresholds for $M = 2$, with a large basin of attraction towards the 'look twice, forgive once' and strict discriminator mix. **c**, Competition between the mix of $M = 2$ and the mixes that evolve in larger $M \in \{4,6,8\}$ in the presence of unconditional strategies. Such dynamics result in new mixed equilibria, which always contain 'look twice, forgive once' and an aggregating discriminator from larger *M*. Specifically, the strict discriminator with $M = 2$ is outcompeted by less strict discriminators but with larger $M \in \{4,6,8\}$. The rates of cooperation at the discriminating equilibria are all >99.5%. For all panels, the benefit-to-cost ratio is $b/c = 5$, with error rates of assessment and execution $\alpha = 0.02$ and $\varepsilon = 0.02$, respectively.





**Extended Data Fig. 10 | Coevolution of strategies and social norms, for all pairs of norms.** Each table cell shows the outcome of the competition among the eight types obtained from pairing one of the four strategies (*ALLC*, *ALLD*, $DISC_{1/2,2}$ and $DISC_{2/2,2}$) with one of the two social norms. The initial proportion of the two social norms is set to 50%. The four different bars in cell represent different probabilities of strategy imitation ($w$) versus social norm imitation ($1-w$). Each bar summarizes steady states reached by numerical integration from 975 uniformly distributed different initial strategy frequencies. Hues represent strategies; the brightness of each hue (lighter or darker) indicates the social norm. The rate of cooperation at the equilibria where the population consists entirely of discriminator strategies exceeds 97.5% in all cases. For all panels, the benefit-to-cost ratio is $b/c = 5$, with error rates of assessment and execution $\alpha = 0.02$ and $\varepsilon = 0.02$, respectively.

nature portfolio

Corresponding author(s): Corina E. Tarnita

Last updated by author(s): Jul 28, 2024

# Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

## Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

| n/a | Confirmed |
|---|---|
| ☒ | ☐ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement |
| ☒ | ☐ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly |
| ☒ | ☐ The statistical test(s) used AND whether they are one- or two-sided<br>*Only common tests should be described solely by name; describe more complex techniques in the Methods section.* |
| ☒ | ☐ A description of all covariates tested |
| ☒ | ☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons |
| ☒ | ☐ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals) |
| ☒ | ☐ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted<br>*Give P values as exact values whenever suitable.* |
| ☒ | ☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings |
| ☒ | ☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes |
| ☒ | ☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated |

*Our web collection on statistics for biologists contains articles on many of the points above.*

## Software and code

Policy information about availability of computer code

| | |
|---|---|
| Data collection | Custom code in Julia v1.8.5 to generate the time trajectories and steady states of the evolutionary dynamics of different sets of strategies. Custom code in Mathematica v14.0 to plot such trajectories. |
| Data analysis | Custom code in Julia v1.8.5. Code DOI [10.5281/zenodo.12795781] and available at the URL: https://github.com/michel-mata/IRMO.jl |

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

## Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

There are no empirical data associated with this study. All synthetic data generated are available at the URL: https://github.com/michel-mata/IRMO.jl/

## Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

| | |
|---|---|
| Reporting on sex and gender | N/A |
| Reporting on race, ethnicity, or other socially relevant groupings | N/A |
| Population characteristics | N/A |
| Recruitment | N/A |
| Ethics oversight | N/A |

Note that full information on the approval of the study protocol must also be provided in the manuscript.

# Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☐ Behavioural & social sciences ☒ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

# Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

| | |
|---|---|
| Study description | Theoretical model and numerical analysis and simulations. |
| Research sample | N/A |
| Sampling strategy | N/A |
| Data collection | Data generated by custom code in Julia v1.8.5 |
| Timing and spatial scale | N/A |
| Data exclusions | N/A |
| Reproducibility | Custom code in Julia v1.8.5. Code DOI [10.5281/zenodo.12795781] and available at the URL: https://github.com/michel-mata/IRMO.jl |
| Randomization | N/A |
| Blinding | N/A |

Did the study involve field work? ☐ Yes ☒ No

# Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

## Materials & experimental systems

| n/a | Involved in the study |
|---|---|
| ☒ | ☐ Antibodies |
| ☒ | ☐ Eukaryotic cell lines |
| ☒ | ☐ Palaeontology and archaeology |
| ☒ | ☐ Animals and other organisms |
| ☒ | ☐ Clinical data |
| ☒ | ☐ Dual use research of concern |
| ☒ | ☐ Plants |

## Methods

| n/a | Involved in the study |
|---|---|
| ☒ | ☐ ChIP-seq |
| ☒ | ☐ Flow cytometry |
| ☒ | ☐ MRI-based neuroimaging |

# Plants

| | |
|---|---|
| Seed stocks | N/A |
| Novel plant genotypes | N/A |
| Authentication | N/A |