A House Divided: Cooperation, Polarization, and the Power of Reputation

Sebastian Wild[1], Phillip Keldenich[2], Jann Spiess[3], Maximilian Schlund[4], Jano Costard[5], Jonas Radbruch[6], Paul Stursberg[7], Sándor P. Fekete[2*]

[1] Department of Computer Science, University of Liverpool, Liverpool L69 3BX, UK.

[2] Department of Computer Science, TU Braunschweig, 38106 Braunschweig, Germany.

[3] Graduate School of Business, Stanford University, Stanford, CA 94305, USA.

[4] School of Computation, Information and Technology, Technical University Munich, 85748 Garching, Germany.

[5] Federal Agency for Disruptive Innovation - SPRIND, 04103 Leipzig, Germany.

[6] School of Business and Economics, Humboldt University Berlin, 10099 Berlin, Germany.

[7] Department of Mathematics, Technical University Munich, 85748 München, Germany.

[*] Corresponding author: s.fekete@tu-bs.de

Altruistic cooperation has enabled humans to thrive[1]. However, the interaction of sentient individuals faces the dilemma of limiting the downsides of personally beneficial, but globally detrimental selfish behavior without causing even more damage through escalating conflicts. The evolution of cooperation has been studied in non-zero sum games, with the *Prisoner's Dilemma,* "the E. coli of social psychology"[2], providing a fundamental test case. Typically[3-12], interactions between individuals may (i) occur repeatedly, (ii) involve groups of individuals, (iii) be subject to evolutionary

mechanisms, often based on the study of equilibria for homogeneous settings.[13] However, a better understanding of the *non-equilibrium dynamics* of cooperation *in structured environments* is crucial for further progress. Here we consider an inhomogeneous, spatial, dynamic setting, in which evolution occurs not necessarily at an equilibrium. We demonstrate how minimal, publicly observable information on previous behavior can be exploited to outperform alternatives, achieving evolutionary performance similar to clandestine, membership-based strategies. We also show how *polarization* (with a cooperating population disintegrating into competing factions) and *tribalism* (with cooperation solely based on group membership instead of behavior) can arise, how these phenomena can be overcome with two additional mechanisms, and how cooperation can erode. Our results demonstrate how cooperation, reputation, polarization and tribalism are intricately linked, even in a simple mathematical model in which they arise in absence of complex psychological mechanisms. This provides a fundamental explanation for how robust cooperation may break down when faced with eroding universality of globally recognized values and of local, direct reciprocity; it may also help to prevent behavior-based reputation systems from giving way to emergent polarization and, ultimately, purely membership-based tribalism. We also anticipate that our methods will be of critical importance for the design and implementation of artificial structures based on the interaction of many independent, self-interested virtual agents.

Classic work[2-13] on the evolution of cooperation analyzes group scenarios with large populations of interacting agents (subject to the aspects (i)–(iii) stated above), introducing evolutionary mechanisms based on superior payoff (in settings such as the Prisoner's Dilemma) for groups of cooperating individuals following a joint strategy. Typically, well-mixed populations are considered, in which all pairs of individuals can interact, so that the evolutionary outcomes are global equilibria. Among the considered mechanisms[3-12,14] are direct and indirect reciprocity, often making use of an evaluation of observable behavior, as well as additional, hidden information.

In contrast, we consider the setting in which interaction (iv) takes place in spatially structured environments. This reflects the fact that evolutionary successful, cooperative groups are often first established locally[15], making it natural to consider populations that interact *spatially*[3,15-21]; see Fig. 1b for a basic model of such spatial interaction. This does not only occur in cell biology, but even in human populations: the impact on partisan sorting of humans was recently highlighted[22]. Other recent work on polarization has considered cognitive aspects[23], and algorithmic complexity[24]. The interplay of reputation and polarization has also been considered[25], but only based on indirect reputation evaluation, which differs from our more general approach; moreover, the resulting form of polarization is different from what we consider here, and inherently unstable.

At the higher level, the competition between different subpopulations with the same respective strategy is based on the success of locally competing individuals, according to their respective payoffs; as a consequence, global success is not necessarily based on centrally coordinated global welfare maximization, but through the dynamic, distributed process of local optimization. This makes it natural to extend the notion of evolutionary success by gauging it with the quantitative parameter of *invasion speed*; refer to Fig. 1 c-f and the Methods section for details of the ensuing spatial dynamics. This goes beyond the traditional notion of an *evolutionary stable strategy* (ESS), which requires only stability against a small fraction of mutant strategies[23] in well-mixed populations without spatial considerations. Instead, we consider how such a stable population can *evolve* in the first place by invading and defeating an existing, well-established population in a spatial setting with localized neighborhoods. Previous work[16,17,21] on the spatial Prisoner's Dilemma shows how unconditional cooperators are able to invade a population of defectors and maintain spatially coherent clusters for mildly adversarial environments with only weak benefit of defecting; however, they quickly die out in more hostile settings.

These limitations of cooperative strategies can be addressed by enhancing them with principles such as indirect reciprocity through the use of natural, publicly available information and algorithmic mechanisms. This involves evaluating observed behavior between individual interactions to gauge trustworthiness, leading to a reputation that is assigned to players[7-10,12,14-15,27-35]. The basic idea is that in large populations, it is possible to observe and learn from the behavior of an individual towards many others, even if the number of interactions between the same two individuals is limited: "Indirect reciprocity describes the interaction between a donor and a recipient. The donor can either cooperate or defect. The basic idea of indirect reciprocity is that cooperation increases one's own reputation, while defection reduces it. The fundamental question is whether natural selection can lead to strategies that base their decision to cooperate (at least to some extent) on the reputation of the recipient." (Nowak[3], supporting online material.) This establishes a setting in which "Each player has an image score, s, which is known to every other player." (Nowak and Sigmund[7]); "An individual's score is known by all group members, for instance because all interactions are publicly observed" (Leimar and Hammerstein[10]). More technically, Nowak and Sigmund[9] noted, "This review of theoretical and empirical studies of indirect reciprocity stresses the importance of monitoring not only partners in continuing interactions but also all individuals within the social network. Indirect reciprocity requires information storage and transfer as well as strategic thinking and has a pivotal role in the evolution of collaboration and communication."

Technically, reputation is captured by a mathematical function that uses a spectrum of information on a player (in particular, observed past actions) as input to compute a decision on cooperation or defection when interacting with that player. This expresses how a player can map an opponent's (potentially extensive) sequence of past decisions to an eventual binary decision of trustworthiness, i.e., an action from {cooperate, defect} in a new interaction. To achieve evolutionary success, a considerable variety of functions have been proposed. In some settings, simple reputation systems may suffice[7,8], but often they are not successful under all circumstances[10,30]. More advanced reputation systems are often able to overcome shortcomings of simpler ones[14,30-32], frequently at the expense of using a larger amount of interaction data. Another option is to use additional, hidden information, such as membership in a clandestine organization: the strategy MAFIA is characterized by a secret bit that is only visible to other members. Both aspects encounter limitations: keeping track of vast amounts of interaction data quickly becomes prohibitively costly, and mechanisms that are based on covert coordination or group membership may be undesirable for other reasons, e.g., in the context of organized crime or racism. Further details are discussed in the Methods section.

We have developed a simple yet powerful mechanism that uses only a minimal amount of *publicly* visible information. Our strategy GANDHI assigns a reputation value of good or bad (corresponding to worthy or unworthy of cooperation, i.e., the actions {cooperate, defect} in the next interaction) to each individual, and conducts updates only based on two bits of information, corresponding to two past interactions with others: An individual is considered good if both (i) its last interaction with another good individual was cooperation, and (ii) its last interaction with a bad individual was non-cooperation. The key idea behind this strategy is to efficiently promote both desired cooperation with trustworthy individuals and punish undesired support of defectors; the name alludes to a well-known quote by Mahatma Gandhi[36]: “Non-cooperation with evil is as much a duty as is cooperation with good.“

We have demonstrated the power of this simple strategy in a systematic comparison with a spectrum of other methods for indirect reciprocity that have been proposed in the literature. To this end, we used extensive multi-parameter computer simulations of a standard model from evolutionary game theory (Fig. 1 a-b), complemented with mathematical analysis of Markov chain approximations. This model isolates the features of a social dilemma in which individuals have no immediate incentive to cooperate; additional mechanisms, in particular, public reputation, can help cooperative strategies gain foothold even when the temptation of defecting is very high. We have followed previous work[15-21,29-30,32-35] on the spatial Prisoner’s Dilemma, which considered a setting in which reputation-based strategies had to attempt invading a population of unconditional defectors (ALLD, which never cooperate) or a population of unconditional cooperators (ALLC, which always cooperate); see Fig. 1. In the context of this spatial version of the classic Prisoner’s Dilemma, GANDHI is never weaker than other reputation-based strategies, and outperforms all of them in terms of the memory required for this success. Furthermore, the performance of GANDHI is comparable to MAFIA, which has access to hidden information; these conclusions are validated by both a detailed mathematical analysis and extensive simulations. Our findings allow us to combine and extend results from previous spatial and indirect reciprocity approaches to scenarios in which cooperation is more costly; our results also contribute to explaining how reputation may have evolved by showing that even a small group of individuals can dominate a large population even in rather adversarial scenarios, establishing reputation as a global mechanism through the course of gradual evolution.

While this demonstrates the power of GANDHI in competition with other strategies, its simple mechanism has one significant downside, which can lead to polarization of an otherwise cooperating population: it is *antisymmetric*, i.e., it remains consistent when good and bad reputation are *swapped*. As shown in Fig. 2 (and discussed in more detail in the SI), this can lead to fragmenting the successful GANDHI population into two competing factions (“red” and “blue”) that both follow GANDHI, despite implementing the exact same set of rules: while members of RED GANDHI (RG) consider other RG individuals as good, but members of BLUE GANDHI (BG) as bad, members of BG consider other BG individuals as good, but members of RG as bad. Such a split can be induced by an inhomogeneous initialization (corresponding, e.g., to optimistic or pessimistic individuals), but may be triggered even by a seemingly innocuous disturbance, such as a single error in observation. Once this fragmentation happens, further observations only strengthen the respective assignments, as BG will cooperate with BG, but not with RG, and vice versa, confirming the respective (but antisymmetric) labeling as good and bad. Even an ongoing competition between the two factions (with individuals being “turned” when overpowered by their neighbors of the other faction) only leads to stronger polarization in which the spatial separation between factions increases, resembling the process of coarsening of spin glasses from physics[37]. This very gradual reduction in separation length (corresponding to the occurrence of non-cooperation between the factions) still manages to slightly improve global welfare (corresponding to overall average score) based on local competition, and thus still outperforms other strategies. However, neither party can decisively defeat the other; moreover, the coarsening process itself proceeds extremely slowly, when compared to the relatively swift evolutionary of success of GANDHI against other strategies. As it turns out, this becomes completely analogous to a contest between two purely membership-based strategies, in which members of RED MAFIA (RM) only cooperate with other members of RM, while members of BLUE

MAFIA (BM) only cooperate with each other. In effect, *polarization* (in which two factions emerge that start to fight each other, based on observable behavior) becomes indistinguishable from *tribalism*, in which cooperation and non-cooperation are not based on behavior, but on group membership alone.

As a consequence, we studied additional mechanisms to deal with polarization. The first is a *global* mechanism based on universally recognized authorities, in which two entities (“virtue” and “evil”) are uniformly considered as good and bad; players interact with these authorities at random occasions (with a probability of h), allowing their reputation to be updated to good even in the perception of the other faction. While this does overcome polarization for sufficiently large values of h, the critical threshold (around $h = 0.735$) appears too high for an effective mechanism by itself.

A second enhancement is a *local* mechanism based on direct reciprocity, in which a direct neighbor is considered good as long as their last direct interaction was to cooperate. In the long run, this can lead to more wide-spread cooperation, but it does not overcome the polarization of reputation, leaving the whole population vulnerable to a collapse of cooperation when local reciprocity is weakened.

However, using both of the global and the local mechanisms[1] in combination with GANDHI (resulting in GANDHI++) is able to counter polarization: for very small values of h, the combination of direct, local reciprocity with global calibration by universally recognized authorities is able to overcome even settings with artificially enhanced polarization (Fig. 3). As a consequence, GANDHI++ is also quite robust against perturbations, making it a very powerful strategy that uses only minimal information and mechanisms.

[1] This may be mapped to the classical “Love God and love your neighbor.”, Matthew 22:36-40 and Mark 12:30-31.

While this sequence of insights is rather encouraging for the development of cooperative mechanisms, a further twist and caveat arises from considering a direct competition between GANDHI++, GANDHI and MAFIA: While GANDHI++ is able to both thrive in basic adversarial settings (such as swiftly defeating populations of ALLD) and also to deal with polarization—thereby achieving universal cooperation—it is vulnerable to populations without the stabilizing effects of globally recognized institutions and local reciprocity, leading to an erosion of cooperation: As we show in Fig. 4, a population of GANDHI++ can slowly but surely be defeated by an opposing group of GANDHI. Furthermore, GANDHI in turns falls prey to MAFIA, due to the slightly slower update mechanism when taking over other players.

Discussion

We are confident that our findings will provide useful tools for the field of systems of artificial agents, where cooperation has to be based on explicitly programmed protocols, and the use and availability of a small amount of publicly available information is of crucial importance. This opens up a number of additional mechanisms and aspects beyond the confines of the considered setting with the Prisoner's Dilemma in a spatial setting with fixed neighborhoods of fixed size; in particular, active mechanisms of expanding connectivity and more variable payoffs in other non-zero-sum games (which allow both group support to "frontier" members faced with adversarial individuals, as well as escalation in conflict) promise further relevant insights for theory and practice.

While we make no claims in the realm of political or social sciences, it seems inevitable that the simplicity of our reputation-based mechanisms makes them particularly suitable to be studied in these important areas. (After all, even a famous quote such as "*A house divided against itself, cannot stand.*" relies on the metaphoric power of gravity.) In particular, it is conceivable that the emergence of increasing tribalism in a society may have some similarities to a transition from Gandhi++ to Gandhi, i.e., the erosion of the polarization-preventing mechanisms of direct reciprocity and universally accepted instances of "virtue" and "evil", which may in turn give way to a transition to the purely membership-based Mafia. Conversely, successfully overcoming tribalism may hinge on (re-)establishing these global and local mechanisms.

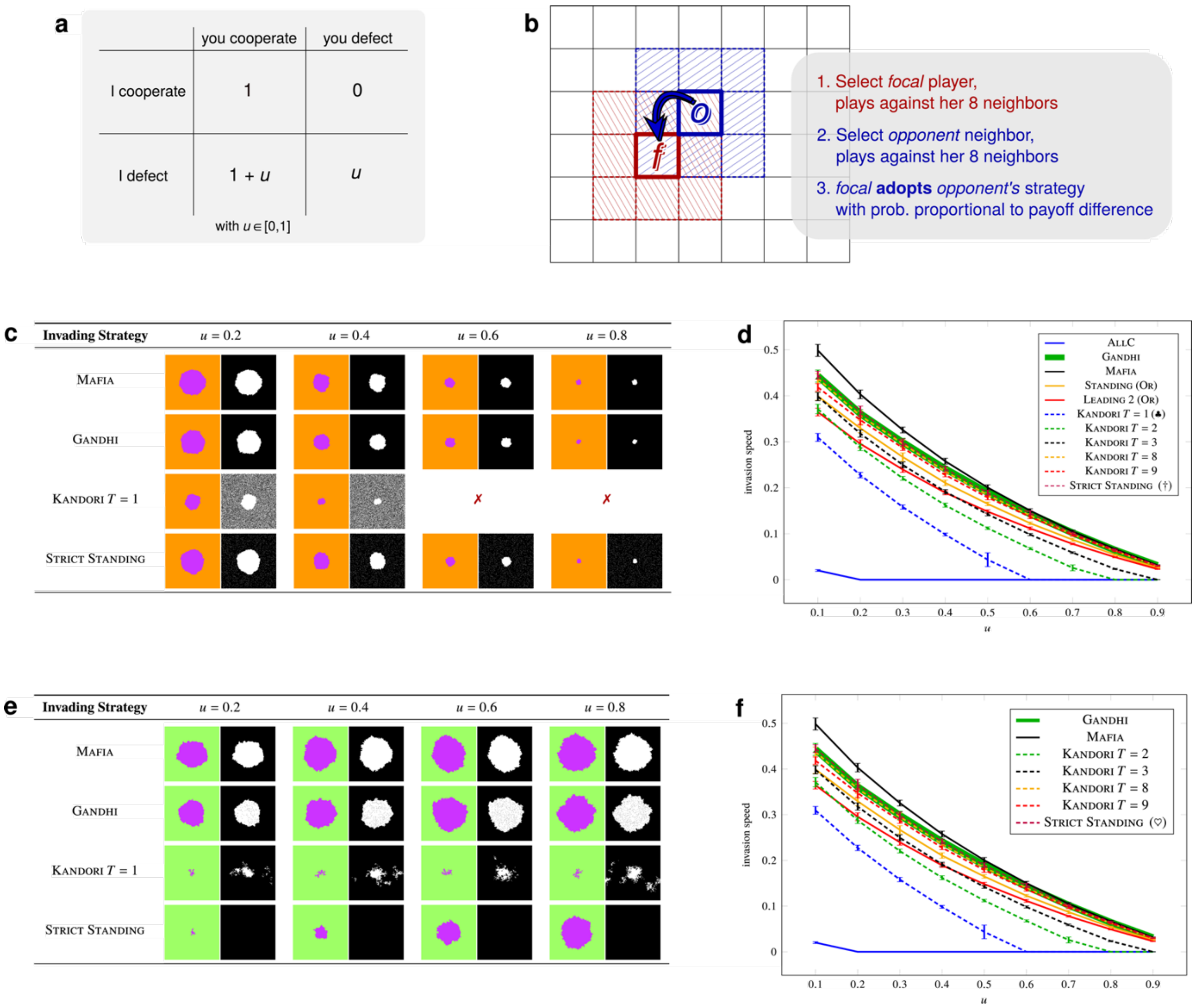


Figure 1: Spatial prisoner's dilemma with semi-deterministic replicator rule and public reputation, and invasion speed of reputation-based strategies (Disc) in an All-Defect (AllD) environment.

a+b, The underlying model[16] with **a,** the payoff matrix for Prisoner's Dilemma (PD) and **b,** repeated player interaction with their eight neighbors on an N × N square lattice of individuals with periodic boundaries, and adopting more successful strategies in a replicator update process. In addition, players have access to public data based on previous interactions. **c**, Prisoner's Dilemma with ALLD (orange) and DISC invaders (purple) on a 200 × 200 grid, and snapshots after 240 generations. Colored tiles: population, B/W tiles: reputation (white = good, black = bad). Rows vary the reputation systems, columns the exploitation benefit u. The reputation system KANDORI with $T = 1$ dies out in an ALLD population for $u \geq 0.6$, demonstrating the weakness of single-bit tracking. **d**, Invasion speed of DISC against an ALLD population under all reputation systems and different exploitation conditions; higher speed is stronger. Each data point shows mean and standard deviation of 20 independent runs. "STRICT STANDING (†)" represents STANDING and STRICT STANDING reputations, "KANDORI $T = 1$ (♣)" stands for KANDORI with $T = 1$, and LEADING 3, 4 and 5. **e**, **f**, Same as **a, b,** but in an ALLC environment (shown as light green). KANDORI reputation does not die out against cooperators, but fails to convert them effectively, leading to fractal structures in the strategies distribution.

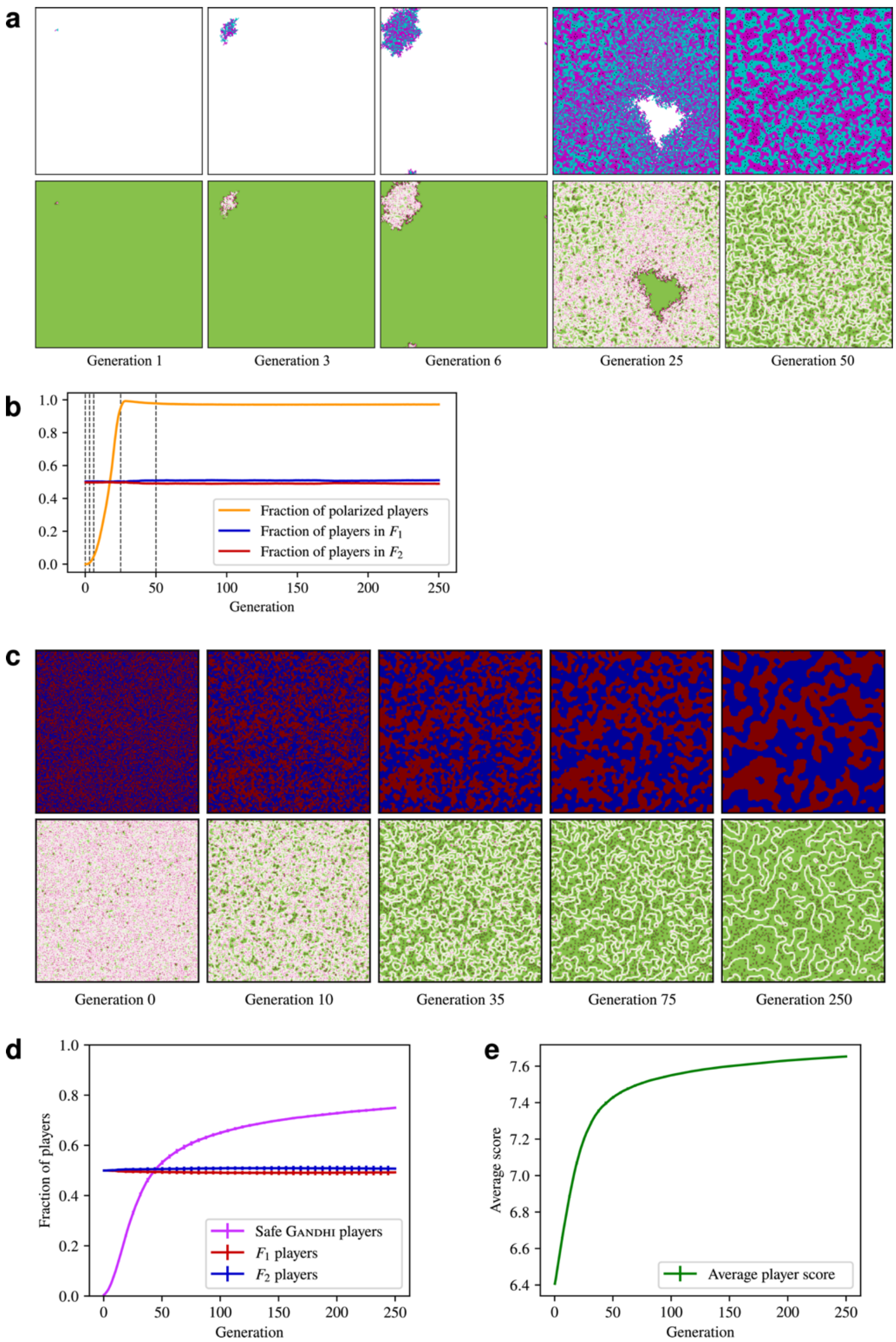


Figure 2: Polarization emerges among two symmetric GANDHI factions. a-b, Polarization, i.e., players seen as good by one faction and bad by the other, spreads from a single misinterpreted duel (in the top left corner). **a**, simulation after 1, 3, 6, 25, and 50 generations; top tiles: reputation difference, bottom tiles: score. In the reputation-difference map, players are cyan if considered good by RED GANDHI (RG) and bad by BLUE GANDHI (BG), magenta if considered bad by RG and good by BG, and black (resp. white) if considered bad (resp. good) by both factions. The score shows the payoff each player

achieved in their last game (greener is better). **b**, Number of polarized players over time. A very small number of players become depolarized; such a player is seen as bad by both factions, because they were the last in a neighborhood to change faction and were hence unable to defect against a bad opponent to regain good reputation with their own faction. **c-e**, Two competing groups of GANDHI, red and blue, over time. **c**, Snapshots of the simulation; top tiles: population, bottom tiles: score. **d**, The number of "safe" players, i.e., players for which all neighbors are in their own faction averaged over n=10 experiments with random initial configurations. This number grows over time through "coarsening" of the boundaries. **e**, Social welfare (average total score that each player gets when playing with their neighbors) over time. This rises in line with safe players, but does not overcome the polarization of the overall population.

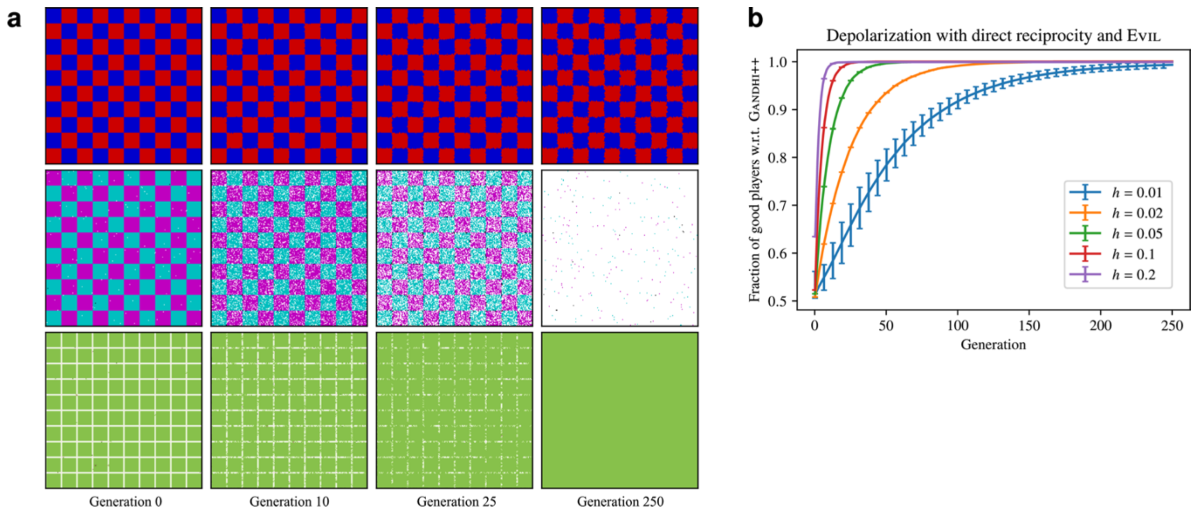


Figure 3: Opposing GANDHI++ factions recovering from initially prevalent polarization. a, Snapshots of a typical simulation on a 200 × 200 grid after 0, 10, 25, and 250 generations. Top tiles: population (red and blue GANDHI factions), middle tiles: reputation difference (colors as in Fig. 3), bottom tiles: scores. Contact probability with `virtue` and `evil` is h = 0.01. Reciprocity and regular contact with global authorities eventually leads to all players being considered `good` by both factions and thus to global cooperation. **b**, Number of polarized players over time for different global-authority probabilities h. Here, only `evil` authorities are used, showing that `virtue` is not necessary in GANDHI++.

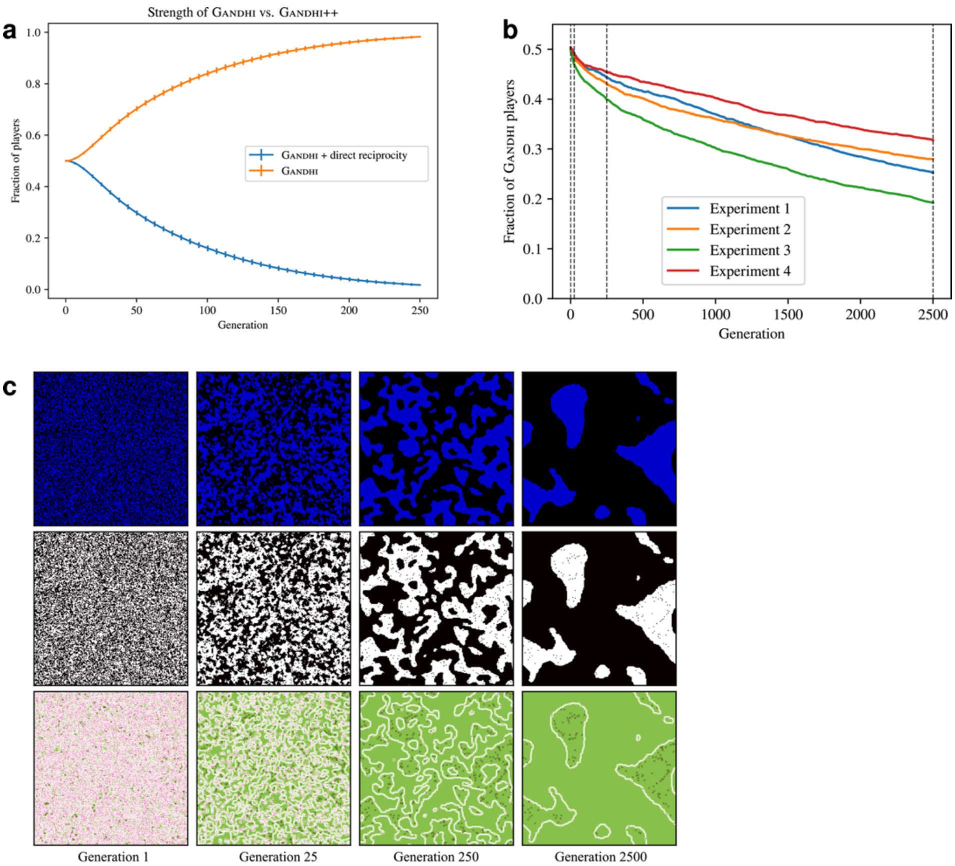


Figure 4: Direct competition of GANDHI++, GANDHI and MAFIA. a, Number of GANDHI and GANDHI++ players over time in the simulation of a direct competition. GANDHI is able to replace GANDHI++ relatively quickly. **b-c**, Direct competition of MAFIA (black) and GANDHI (blue). **b**, The number of GANDHI players over time for four exemplary simulations. **c**, Snapshots from Simulation 1; top tiles: population, middle tiles: GANDHI's reputation, bottom tiles: scores. Similar to a competition between two MAFIA or two GANDHI factions, we observe a coarsening of the strategy distribution. MAFIA eventually overcomes GANDHI, but the process is orders of magnitude slower. Only few GANDHI players on the boundary of the resulting large blocks of GANDHI players are vulnerable.

References and Notes:

1. Tomasello, M., Melis, A. P., Tennie, C., Wyman, E. & Herrmann, E. Two Key Steps in the Evolution of Human Cooperation: The Interdependence Hypothesis. *Current Anthropology,* **53,** 673-692 (2012).
2. Axelrod, R. Effective choice in the Prisoner's Dilemma. *Journal of Conflict Resolution* **24**, 3–25 (1980).
3. Axelrod, R. & W. D. Hamilton, W. D. The evolution of cooperation. *Science* **211**, 1390–1396 (1981).
4. Nowak, M. A. Five rules for the evolution of cooperation. *Science* **314**, 1560–1563 (2006).
5. Nowak, M. A. Evolving cooperation. *Journal of Theoretical Biology* **299**, 1–8 (2012).
6. Doebeli, M. & Hauert, C. Models of cooperation based on the prisoner's dilemma and the snowdrift game. *Ecology Letters* **8**, 748–766 (2005).
7. Nowak, M. & Sigmund, K. Evolution of indirect reciprocity by image scoring. *Nature* **393**, 573–577 (1998).
8. Nowak, M. A. & Sigmund, K. The dynamics of indirect reciprocity. *Journal of Theoretical Biology*, **194**, 561–574 (1998).
9. Nowak, M. A. & Sigmund, K. Evolution of indirect reciprocity. *Nature* **437**, 1291–1298 (2005).

10. Leimar, O. & Hammerstein, P. Evolution of cooperation through indirect reciprocity. *Proceedings of the Royal Society of London. Series B: Biological Sciences* **268**, 745–753 (2001).

11. Imhof, L. A., Fudenberg, D. & Nowak, M. A. Evolutionary cycles of cooperation and defection. *PNAS* **102**, 10797-800 (2005).

12. Sigmund, K. Moral assessment in indirect reciprocity. *Journal of Theoretical Biology* **299**, 25–30 (2012).

13. Nash, J. F. Equilibrium points in n-person games, *PNAS* **36**, 48-49 (1950).

14. Ohtsuki, H. & Iwasa, Y. How should we define goodness? — Reputation dynamics in indirect reciprocity. *Journal of Theoretical Biology* **231**, 107–120 (2004).

15. Helbing, D., Szolnoki, A. Perc, M. & Szabó, G. Evolutionary establishment of moral and double moral standards through spatial interactions. *PLoS Computational Biology* **6**, e1000758 (2010).

16. Fu, F., Nowak, M. A. & Hauert, C. Invasion and expansion of cooperators in lattice populations: Prisoner's dilemma vs. snowdrift games. *Journal of Theoretical Biology* **266**, 358–366 (2010).

17. Hauert, C., Spatial effects in social dilemmas. *Journal of Theoretical Biology* **240**, 627–636 (2006).

18. Szabó, G. & Tőke, C. Evolutionary prisoner's dilemma game on a square lattice. *Physical Review E* **58**, 69–73 (1998).

19. Nowak, M. A. & May, R. M., Evolutionary games and spatial chaos. *Nature* **359,** 826–829 (1992).

20. Hauert, C. & Doebeli, M. Spatial structure often inhibits the evolution of cooperation in the snowdrift game. *Nature* **428**, 643–646 (2004).

21. Langer, P., Nowak, M. A. & Hauert, C. Spatial invasion of cooperation. *Journal of Theoretical Biology* **250**, 634–641 (2008).

22. Brown, J. R. & Enos, R. D. The measurement of partisan sorting for 180 million voters. *Nature Human Behaviour* 1–11 (2021).

23. Wu, J. S.-T. Wu, Hauert, C., Kremen, C., & Zhao, J. A framework on polarization, cognitive inflexibility, and rigid cognitive specialization. *Frontiers in Psychology* 13, Article 776891 (2022).

24. Chatterjee, K., Ibsen-Jensen, R. Jecker, I., & Svoboda, J. Complexity of spatial games. *42nd IARCS Annual Conference on Foundations of Software Technology and Theoretical Computer Science (FSTTCS),* 11:1--11:14 (2022).

25. Gross, J., & De Dreu, C.K.W. The rise and fall of cooperation through reputation and group polarization. *Nature Communications* 10, 776 (2019).

26. Maynard Smith, J. & Price, G. R., The logic of animal conflict. *Nature* **246**, 15–18 (1973).

27. Sugden, R. The economics of rights, co-operation and welfare, B.Blackwell, 1986.

28. Kandori, M. Social norms and community enforcement. *The Review of Economic Studies* **59**, 63–80 (1992).

29. Nakamaru, M. & Kawata, M. Evolution of rumours that discriminate lying defectors. *Evolutionary Ecology Research* **6**, 261–283 (2004).

30. Panchanathan, K. & Boyd, R. A tale of two defectors: the importance of standing for evolution of indirect reciprocity. *Journal of Theoretical Biology* **224**, 115– 126 (2003).

31. Sigmund, K. & Brandt, H. The logic of reprobation: assessment and action rules for indirect reciprocation. *Journal of Theoretical Biology* **231**, 475–486 (2004).

32. Panchanathan, K. Two wrongs don't make a right: The initial viability of different assessment rules in the evolution of indirect reciprocity. *Journal of Theoretical Biology* **277**, 48–54 (2011).

33. Brandt, H. & Sigmund, K. The good, the bad and the discriminator — errors in direct and indirect reciprocity. *Journal of Theoretical Biology* **239**, 183–194 (2006).

34. Yang, W., Juan W., & Chengyi X. "Evolution of cooperation in the spatial public goods game with the third-order reputation evaluation." Physics Letters A 383.26 (2019): 125826.

35. Quan, J., Qin, Y., Zhou, Y., Wang, X., & Yang, J. B. (2020). How to evaluate one's behavior toward "bad" individuals? Exploring good social norms in promoting cooperation in spatial public goods games. *Journal of Statistical Mechanics: Theory and Experiment*, 2020(9), 093405.

36. Gandhi, M., Statement before Mr. C. N. Broomfield, I. C. S., District and Sessions Judge, Ahmedabad, 18 March, 1922.

37. Matsuda, H. N., Ogita, A., Sasaki, A. & Sato, K. Statistical mechanics of population: The lattice Lotka-Volterra model. *Progress in Theoretical Physics* **88**, 1035-1049 (1992).

Methods

The overall model in the main text is based on three different aspects: Player actions and interactions lead to payoffs according to a *non-zero-sum game-theoretic setting* (Fig. 1a), the relative success of different strategies leads to their spread in a *spatial setting* (Fig. 1b); choosing cooperation or non-cooperation when interacting with another player can be based on information from previous actions, condensed in *reputation systems*. Here we introduce the main technical aspects; a full account of ensuing data is provided in the Supplementary Information.

Spatial Replicator Dynamics

Interaction between players occurs in a setting in which the global population is structured in local environments; this corresponds to a spatial setting with geometric neighborhoods. We make use of the model by Fu et al.[16], who consider an $N \times N$ square lattice of individuals with periodic boundaries, in which two players repeatedly interact with their eight neighbors by playing a symmetric 2×2 game, as shown in Fig. 1b. To evaluate the evolutionary success of different strategies, we model their spread by using the replicator rule (called semi-deterministic updating by Fu et al.[16]): We randomly choose one focal player out of the $N \times N$ square lattice and an opponent among its eight neighbors. (Using randomly selected duels for potential updates avoids artifacts of synchronization; for settings with stronger parallelization, the expected values for spread can simply be adjusted to filter out the effects of expected waiting times for a duel to occur.) Both play against all their neighbors, resulting in accumulated payoffs P_f and P_o for focal and opponent, respectively. Then the focal player adopts the opponent's strategy with probability

$$f(P_f, P_o) = \begin{cases} \dfrac{P_o - P_f}{8(1+u)} & P_f < P_o \\ 0 & \text{otherwise.} \end{cases}$$

Here $8(1+u)$ is the maximal payoff difference in Prisoner's Dilemma. The replicator rule can be seen as a way to apply the classic replicator dynamics for infinite well-mixed populations[38] to finite structured populations; in both cases, the spreading rate is linear in the payoff differences and the payoffs are based on the mean (neighboring) opponent player[39].

Finally, focal and opponent's reputation is updated according to the respective reputation system. (The reputation of the other neighbors involved in the duels, i.e., the neighbors of focal and opponent, remains unchanged; modifying this assumption would make reputation-based mechanisms only stronger.) We stress that the reputation update is done irrespective of whether the focal player adopts the opponent's strategy; in particular, her reputation is not newly initialized to some reputation score nor is it copied from the opponent's reputation. The spreading of strategies is a mechanism of learning or imitating behavior; such a strategy change is an internal, hidden event that can only be observed by others through subsequent actions. This reflects a setting in which distinction between individuals and their actions is based on location, not on publicly announced strategies.

Reputation Systems

Keeping track of the trustworthiness of players leads to assigning a reputation to players, i.e., a function that uses a spectrum of information on a player (in particular, observed previous actions) to result in a decision on cooperation or defection when interacting with that player: Every player follows a strategy, which is a function that takes that player's and her opponent's reputation as arguments and returns an action from {`cooperate`, `defect`}. The simplest strategies are the unconditional cooperators (ALLC) and unconditional defectors

(ALLD), which do not make use of any reputation; more sophisticated are discriminating (DISC) strategies, which cooperate if the opponent has a good reputation and defect otherwise. The meaning of label good depends on the specific reputation system (such as GANDHI). Because the action space is binary, it suffices to consider only binary reputation values, i.e., players are always either good or bad in the eyes of a DISC strategy. Note that even though the labels good and bad may seem to suggest a moral verdict, our setting does not a priori reward conformal behavior.

In our base model, we assume that all players have the same information as their neighbors, modeling a well-connected world with rapid information dissemination and perfectly observable actions; as discussed further down, there may be additional, hidden information. Different discriminating strategies can use different rules to assign reputation labels and may come to a different verdict based on (the same) past behavior. Formally, a reputation system determines a label good or bad for each individual, based on the history of interactions the individual was involved in. It is important to note that we allow for the possibility of including the reputation of former opponents as well, i.e., players have access to higher-order information. For example, the reputation system may rate defection against good or bad players differently. In the base model, we assume that an individual's reputation is globally agreed upon and based on public information. To model the equivalence in the parallel interaction with all neighbors, we update reputation only after all eight neighbor duels of one propagation round have taken place. This also accounts for a delay in the exchange of information between neighbors until more tangible outcomes are visible; more responsive update rules only enhance the advantage of discriminator systems.

In previous work, a wide spectrum of reputation functions have been proposed; these include IMAGE SCORING by Nowak and Sigmund[7,8], which tracks the balance of previous cooperate and defect actions, but is unable to distinguish between defecting from cooperative or non-cooperative players, GOOD STANDING by Sugden[24] and Leimar and Hammerstein[10], which performs one-bit updates, making it unable to sanction cooperation with non-cooperative players, KANDORI[25], which tallies a player's score over T rounds and only cooperates when desirable behavior is maintained (requiring $\lceil \log_2 (T + 1) \rceil$ bits and punishing one-time noncompliance through T rounds), and the LEADING EIGHT of Ohtsuki and Iwasa[14], which are based on various 1-bit updates. All these differ from our strategy GANDHI, which only uses two bits, but achieves better performance, as demonstrated in the sequel.

Success of GANDHI

To compare the discriminatory efficacy of different reputation systems, we study the following questions: (1) Can a cluster of individuals who follow a joint reputation-based strategy convince members of other strategies to imitate their discriminating behavior? (2) If so, how does evolutionary success compare quantitatively, i.e., how fast is this invasion?

In answer to these questions, we provide both qualitative and quantitative evidence that GANDHI outperforms other similar strategies.

Qualitative Evidence

Initially, all players in the $N \times N$ ($N = 70$) grid use the same incumbent strategy (either ALLD or ALLC), except for a 5×5 square cluster of invading DISC individuals in the middle. We explore every possible combination of incumbent strategy and reputation system of the invading DISC players. Moreover, we vary the exploitation surplus parameter $u \in \{0.1, 0.2, \ldots, 0.9\}$. (We have also carried out a large range of additional experiments against mixed populations. These results are not included here, as they do not provide any additional insights.) For each of these setups, the simulation runs until either the invaders die out or the first invader touches the boundary. By the time the boundary is reached, the invasion's final success can be reliably assessed; further progress would be artificially slowed down by boundary effects.

Some examples are shown in Fig. 1 c+e, with a full overview listed in Extended Data Fig. 1; in addition to Prisoner's Dilemma (PD), the latter also include analogous results for Snowdrift (SD), a two-player non-zero sum game in which cooperation with a non-cooperating opponent is less detrimental. It can be seen that only a limited number of strategies succeed in defeating both ALLD and ALLC populations: KANDORI (with at least $T = 8$, i.e., higher-order interaction data), MAFIA (which uses hidden information) and GANDHI.

Quantitative Evidence

Similar to the observations of Fu et al.[16], expansion basically proceeds at constant speed in both dimensions. Therefore, the square root of the number of DISC players grows linearly in the number of played duels. We accordingly define invasion speed as the corresponding rate of change, i.e., by how much the square root of the number of DISC players grows on average in one generation. A generation is here defined as N^2 simulated duels, which corresponds to one chance per player to reproduce on average. As the snapshots in Fig. 1 c+e show, the region occupied by DISC players is of roughly circular shape. Thus, the invasion speed corresponds to the average growth rate of the radius of this circle.

Fig. 1 d+f show the invasion speed of DISC players using various reputation systems. For reputation systems that could never invade, no line is shown. Each point shows the average invasion speed of 20 independent runs of the corresponding simulation. Error bars show one standard deviation around the mean. The narrow error bars show that invasion speed is a robust measure: It is reliably reproduced in independent runs. As invasion speed is a global measure determined from many independent random variables, low variance was to be expected.

Again, GANDHI dominates all other strategies, with the exception of MAFIA, which achieves faster update speed through hidden information.

Mathematical Evidence

Additional mathematical evidence can be obtained by analyzing the behavior of a Markov chain that models the strategy transition of individuals in a mixed population. For the speed ψ_+ of MAFIA vs. ALLD, this yields

$$\psi_+ = \frac{1}{8}\frac{1-u}{1+u}$$

For the analogous case of GANDHI vs. ALLD, we get a speed of

$$T^{-1} = \frac{-5u^2 - 3u + 8}{41u^2 + 111u + 72}$$

which works out to

$$T^{-1} \approx \frac{1}{9}\frac{1-u}{1+u}.$$

See the Supplementary Information for details of this analysis. As Extended Data Fig. 2 shows, this quantitative correspondence is supported by numerical evidence.

Tribalism

The success of MAFIA relies not on a sophisticated strategy, but on strong group coherence, purely based on membership, i.e., *tribalism*. As a consequence, evolutionary success corresponds to the ability of the group to deal with adversarial groups, including other groups that also pursue MAFIA. Remarkably, two different groups of MAFIA cannot overcome each other, but still manage to improve global welfare (corresponding to overall average score) based on local competition. We demonstrate this with a number of experiments; see Extended Data Fig. 3 for an overview. Starting with an initial random distribution of two different group (REDMAFIA and BLUEMAFIA), running the replicator dynamics leads to a process resembling coarsening of spin glasses from physics[35]. More precisely, local competition between the two populations leads to a shortening of the separating boundary, as a weakly connected member of one population will be surrounded by a majority of members of the other; therefore, such an outlier will perform worse than a duel opponent, which is better connected to members of its own group. As a consequence, local majorities will take over their opposing neighbors, leading to smoother, shorter boundaries between the populations, corresponding to improved average score. (Note that this is only the case in the absence of escalation in the interaction with the opposing group.) However, this growing separation and local symmetry also makes it harder to take over neighbors, so that no subpopulation can defeat the other.

Polarization

GANDHI is not based on membership, so it is more open to cooperating with (and thus benefiting from) neighbors, regardless of their strategy. However, its reputation system is subject to antisymmetry in the following sense. Suppose that there are two factions that both play according to GANDHI, with each faction perceiving its own players as good and the players of the other faction as bad. The players of each faction then consistently cooperate with players of their own faction, but defect against players of the other faction. We call such a population *polarized*. As a consequence, the dynamics play out analogously to two MAFIA factions; see Fig. 2. This implies that there is no inherent mechanism in GANDHI to overcome polarization — once a population is polarized, it remains polarized, and only local boundary minimization (and thus, local improvement of average scores) occurs; refer to Extended Data Fig. 4.

There are several possible sources for polarization. Firstly, polarization may stem from differences in initialization: If one ("REDGANDHI") faction F_1 "pessimistically" initializes all players to a bad reputation and another ("BLUEGANDHI") faction F_2 "optimistically" initializes all players to a good reputation, players in F_1 will defect in their first game; similarly, players in F_2 will cooperate. Both actions are perceived as bad by the other faction, leading to polarization. Secondly, even a single misperception can lead to a global polarization, exposing the fragility of non-polarized populations in the base model. Suppose we start with two GANDHI factions F_1 and F_2 sharing the same initialization and then play a duel for which the action of a single player is perceived as cooperate by F_1 and defect by F_2. This results in a single polarized player who is seen as good by one faction and as bad by the other. Starting from this player, polarization spreads with every duel involving unpolarized and polarized players, until the entire population is polarized; see Fig. 2.

Global Authority

Overcoming polarization in GANDHI requires breaking the antisymmetry between any kind of split into REDGANDHI and BLUEGANDHI. One way to achieve this is by introducing global authorities, virtue and evil, that are unequivocally seen as good resp. bad by any player irrespective of their reputation system. In our simulation, we add these as artificial players that focal and opponent encounter with a probability h after playing the 8 duels with their neighbors. The outcome of the (imaginary) duels with virtue and evil are only used in updating a player's reputation; no payoff results from these encounters. As Extended Data Fig. 5 shows, polarization can be dissolved for sufficiently large values of h: If players see a global authority after at least 73% of the duels, polarization vanishes. Below this threshold, some polarized players remain present (Extended Data Fig. 6) and continuously act as seed for new polarization. Extended Data Fig. 5 also demonstrates that the fraction of polarized players remains relatively stable over time. In isolation, global authorities are only an effective cure for polarization if they are nearly omnipresent.

Local Reciprocity

Another mechanism to potentially counter polarization is to complement globally reported information with direct observations, so that sporadic friendly acts among neighbors may be rewarded and perpetuated. We incorporate this in our model in the form of local reciprocity: Each player remembers for her 8 neighbors the last action they played against her, and considers a neighbor p as good whenever p cooperated with her or when p's global reputation is good.

Extended Data Fig. 7 shows that this added leniency allows two polarized factions of RedGandhi and BlueGandhi to cooperate with each other: After an initial period, all players manage to establish trust at the local level and hence benefit from the maximal social welfare that cooperation entails. However, in the global reputation, polarization still looms large: about half of the players are still polarized, and almost all other players are now globally seen as bad (for not defecting from evil). Local reciprocity can effectively stop polarization from affecting actions, but it does not cure the underlying divide.

GANDHI++

The GANDHI++ reputation system consists of simultaneously using global authorities and local reciprocity in GANDHI. Neither of these two additions in isolation cures polarization in GANDHI; however, GANDHI++ not only stops polarization from emerging, but can restore unity in an existing, completely polarized state (Fig. 3a). Note that any positive probability $h > 0$ for encountering global authorities will eventually lead to the eradication of polarization (Fig. 3b). Incidentally, having regular contact to `virtue` is not needed for this; a global `evil` authority (i.e., a universally regarded adversary) is sufficient.

Eroding Cooperation

While GANDHI++ is able to overcome both basic adversarial settings (such as swiftly defeating populations of ALLD) and deal with polarization, thereby achieving universal cooperation, it is also vulnerable to populations without the stabilizing effects of globally recognized institutions and local reciprocity, leading to an erosion of cooperation: As we demonstrate in the following, a population of GANDHI++ can be defeated by an opposing group of GANDHI, which in turns falls prey to Mafia.

We observe that in a direct confrontation, GANDHI++ loses against GANDHI; see Fig. 4a for a typical outcome. This does not change if we remove the global recognition for `evil` and `virtue` by setting $h = 0$, i.e., even though local reciprocity alone does not suffice to overcome polarization, its presence is already sufficient to lose out against unmodified GANDHI; see Fig. 4a. This phenomenon can be attributed to the following mechanisms. Suppose that a GANDHI++ player p has both GANDHI and GANDHI++ neighbors, gets beaten in a duel by some GANDHI player q in the neighborhood, and changes membership to GANDHI. Player p now considers the GANDHI++ neighbors `bad` and defects against them; each GANDHI++

neighbor r will cooperate with p until r itself has been betrayed by p. This makes GANDHI++ more vulnerable to defectors than GANDHI, which learns not to trust p after a single defection against any GANDHI player. This demonstrates the crucial role of both the existence of universally recognized instances of good and bad, as well as local reciprocity. Therefore, protecting cooperation against polarization hinges on protecting these mechanisms.

From Polarization to Tribalism

While GANDHI exhibits similar power against simple-minded strategies (such as ALLD or ALLC), which are defeated almost as swiftly as by MAFIA, it slowly loses out to MAFIA in a direct confrontation. The speed at which this happens can vary considerably, based on random initialization and duel selection; see Fig. 4b. However, the eventual outcome is inevitable, as long as the update speed for GANDHI (which relies on publicly visible reputation information) is slightly slower than for MAFIA (which only needs to update a hidden bit of information), as observed and analyzed above. A remedy to address this could be to delay the adoption of MAFIA membership by exposed individuals.

Additional References:

38. Taylor, P. D., Jonker, L. B. Evolutionary stable strategies and game dynamics. *Mathematical Biosciences*, 40(1-2):145–156, July 1978.
39. Roca, C. P., Cuesta, J. A. & Sánchez, A. Evolutionary game theory: Temporal and spatial effects beyond replicator dynamics. *Physics of Life Reviews*, 6(4):208–49, 2009.

Acknowledgements

We thank Hella-Franziska Hoffmann, Alexander Koch, Dominik Leipold for initial discussions. The roots of this work were laid at a graduate school of the German Academic Scholarship Foundation; we thank Guido Schäfer, Carolin Bernhofer, Lukas Brantner, Nikolai Diekert, Aleksandrs Galickis, Anna-Klara Grosswendt, Rebecca Hartje, Steffen Karalus, Herrmann-Johannes Kerl, Sarah Marzi, Lorenz Pfeifroth, Erik Schaffernicht, Markus Schepke, Jonathan von Schröder, Jörg Spenkuch, Alexandra Surdina, Eric Szperlinsi, Felix Wersich, Bixuan Xu for their participation.

Author contributions

All authors conceived the study, performed the analysis, discussed the results and wrote the manuscript.

Competing interests

The authors declare no competing interests.

Supplementary information

S.I. is available for this paper, and submitted in parallel.

Data Availability

All described data is available upon request and will be posted at a public repository.

Ethics & Inclusion

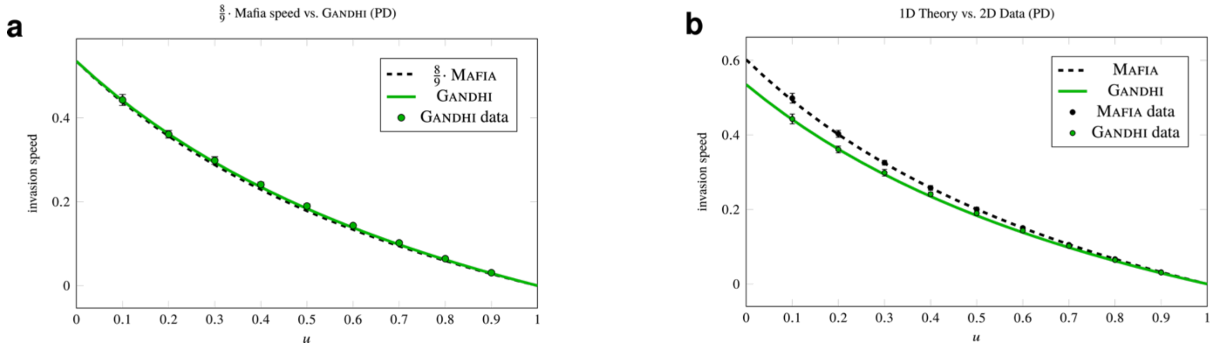
The nature of this work does not involve resource-poor settings.

Correspondence and requests for materials should be addressed to S.P.F.

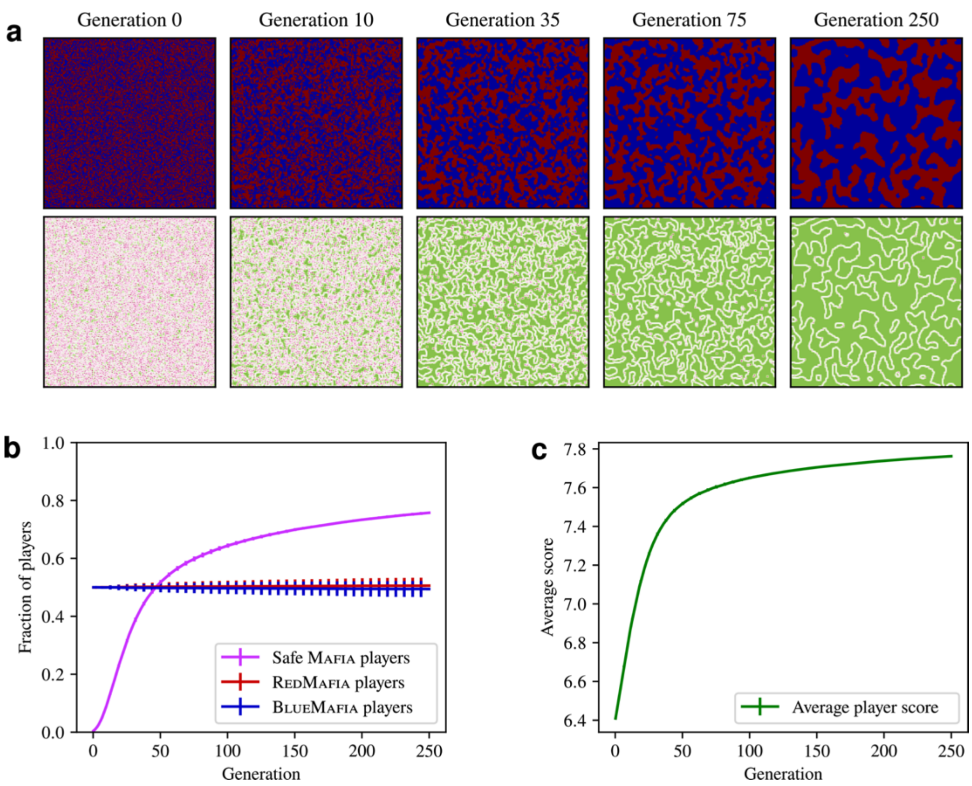
Extended Data Figures

Reputation System	PD AllD	PD AllC	SD AllD	SD AllC
none (AllC)	[0.1, 0.2]	—	[0.7, 0.8]	—
Image Scoring	✗	✗	✗	✗
Strict Standing	✓(†)	✓(♡)	✓(†)	✓(♡)
Standing	✓(†)	✗	✓(†)	✗
Standing (Or)	✓	✗	✓	✗
Leading 2 (Or)	✓	✗	✓	✗
Leading 3	[0.5, 0.6](♣)	✗	✓(♣)	✗
Leading 4	[0.5, 0.6](♣)	✗	✓(♣)	✗
Leading 5	[0.5, 0.6](♣)	✗	✓(♣)	✗
Leading 8	✗	✓(♡)	✗	✓(♡)
Kandori $T = 1$	[0.5, 0.6](♣)	fractal	✓(♣)	fractal
Kandori $T = 2$	[0.7, 0.8]	✓	✓	✓
Kandori $T = 3$	[0.8, 0.9]	✓	✓	✓
Kandori $T = 8$	✓	✓	✓	✓
Kandori $T = 9$	✓	✓	✓	✓
Gandhi	✓	✓	✓	✓
Mafia	✓	✓	✓	✓

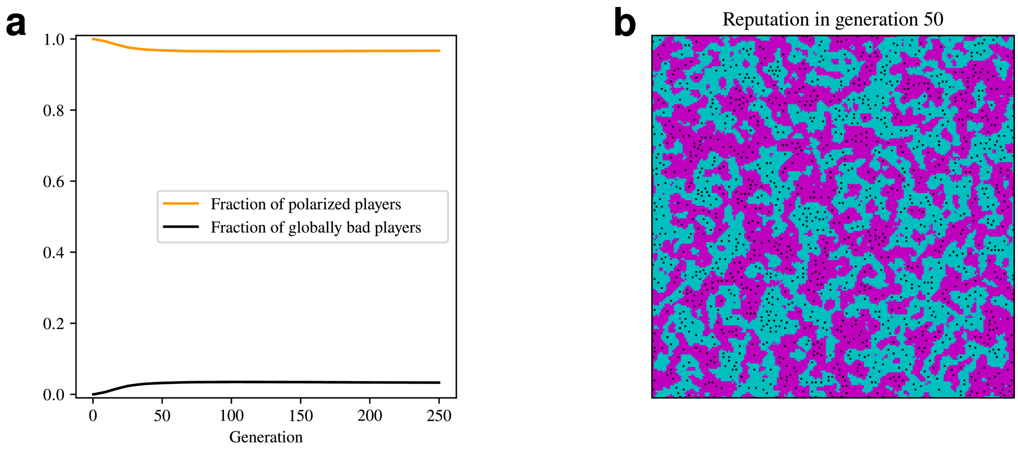
Extended Data Figure 1: Qualitative results on discriminatory efficacy. Each entry shows whether the corresponding reputation system allows Disc to take over the incumbent population in the corresponding setting. Rows with (Or) correspond to scenarios where the Or strategy is used instead of Disc, see Supplementary Information. An entry ✓ means invasion is successful, ✗ means no invasion. An interval $[a, b]$ indicates that invasion depends on the exploitation benefit and the threshold value lies in this interval. The term "fractal" is used when the Disc region forms a fractal-like shape. As only a small fraction of the players joined Disc here, "fractal" counts as ✗. For some settings, several reputation systems become strongly equivalent, i.e., they behave exactly the same in every single step. These equivalence classes are marked by ♥, ♣, and †, respectively.



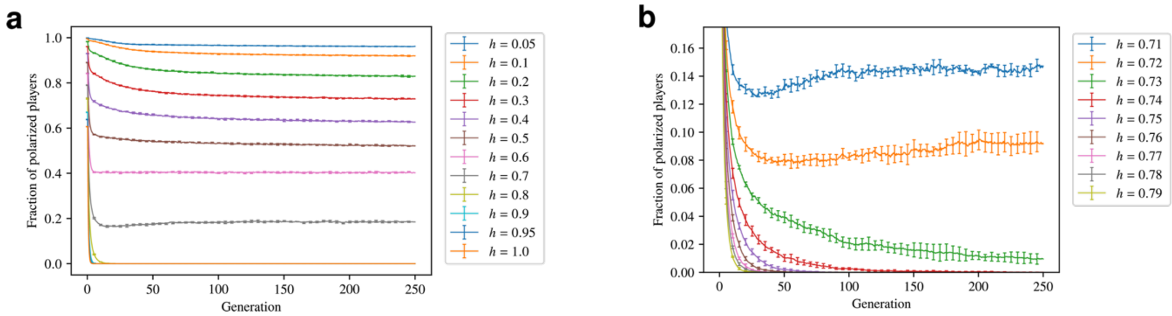
Extended Data Figure 2: Validation of prediction of the one-dimensional Markov model on simulation data. a, The predicted invasion speed of GANDHI (green) from the 1D Markov model for the Prisoner's Dilemma against ALLD (see Supplementary Information) as a function of *u* against 8/9 times the predicted invasion speed of MAFIA (dashed black). **b**, Plot of the (scaled) predicted invasion speed from the 1D Markov model with the actual invasion speed determined from our simulation (as in Fig. 1c+d) for both GANDHI (green) and MAFIA (dashed black). The dependency in *u* matches the theoretical prediction extremely well.



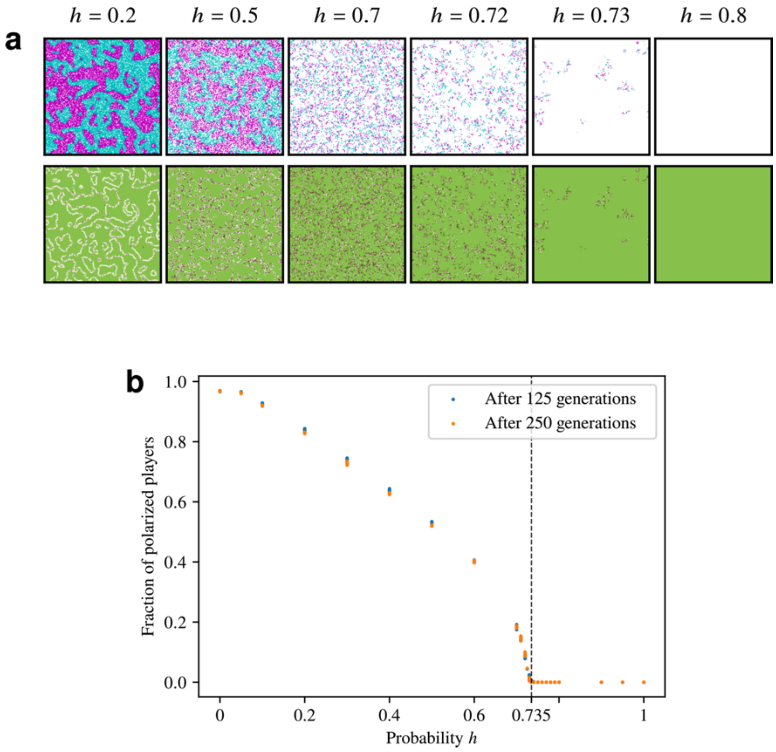
Extended Data Figure 3: Two competing groups of Mafia over time. a, The distribution of strategies (top tiles, red or blue) and the score each player achieved (bottom tiles, greener is better) in the last round they played in an exemplary experiment after the stated number of generations. An initially fine-grained distribution of players, assigned to a group uniformly at random, coarsens over time. **b**, The number of players that are "safe", i.e., completely surrounded by players in their own group (average over n=10 experiments), increases over time in this coarsening process. **c**, The average player score likewise increases over time.



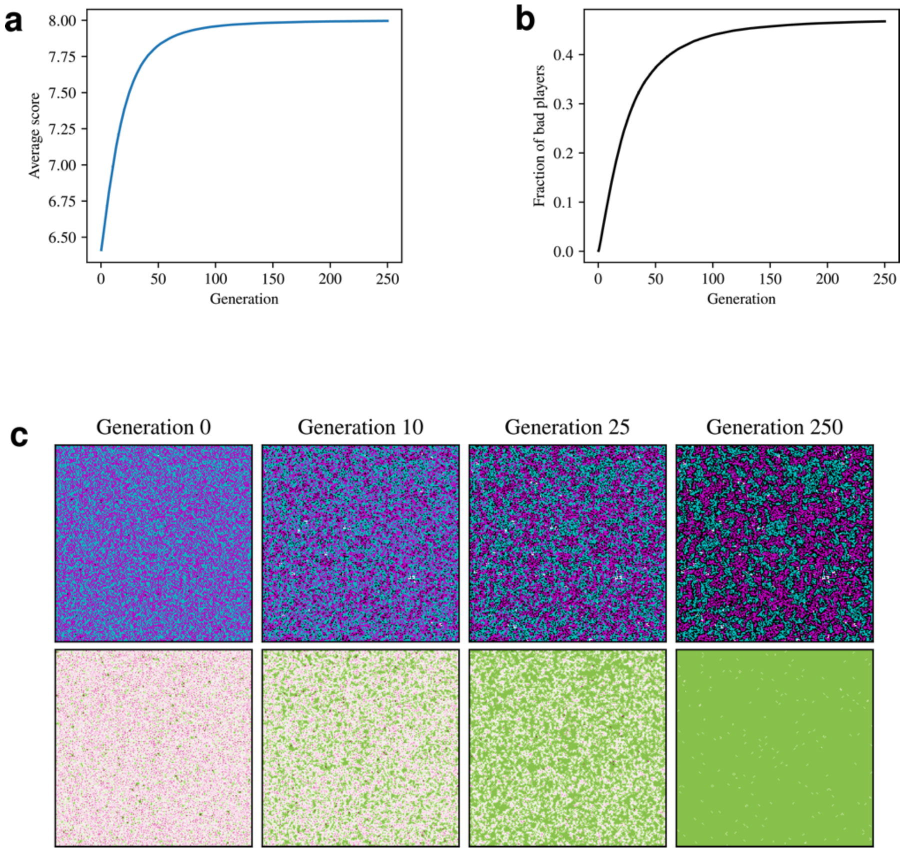
Extended Data Figure 4: Two competing GANDHI factions over time. a, The number of polarized players, i.e., players that are seen as good by one faction and bad by the other, over time, as well as the number of players seen as bad by both GANDHI factions F_1 and F_2. A generation is a number of rounds that corresponds to the total number of players. **b**, The difference in reputation between the two factions. Cyan players are considered good by F_1 and bad by F_2, magenta players are considered bad by F_1 and good by F_2. Black players are seen as bad by both factions. A very small number of players become depolarized, who are now seen as bad. These players are seen as bad by both factions as a result of being the last player in a neighborhood that changed faction — they are unable to defect against a bad opponent to gain good reputation with their own faction because all their neighbors are good. No players are considered good by both factions.



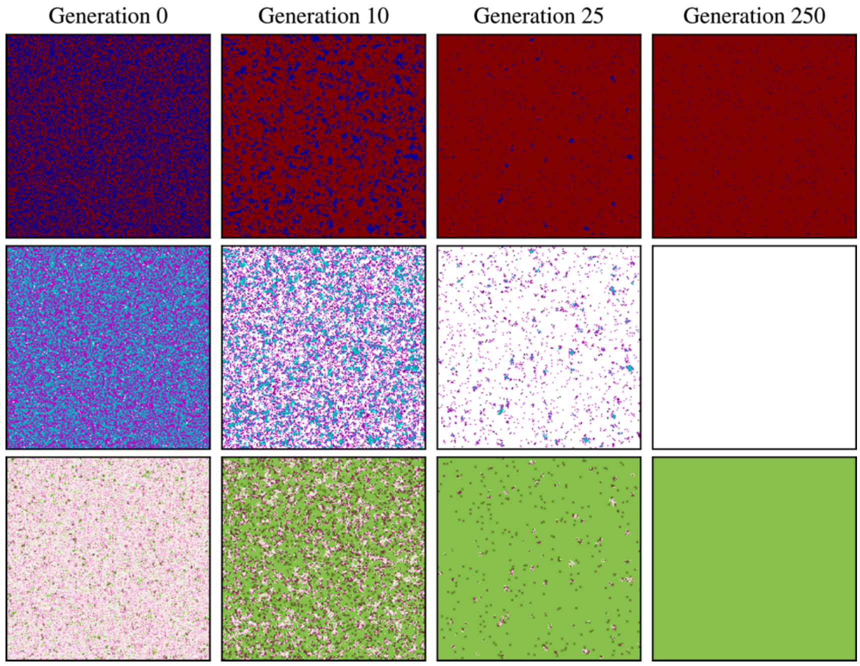
Extended Data Figure 5: The effect of global authorities on the number of polarized players in two competing GANDHI factions. a, b, The number of polarized players over time for two competing GANDHI factions and various values of *h*, the probability of `virtue` and `evil` participating in a duel. Each point is the average of n=10 independent simulations, error bars show one standard deviation. If *h* is not high enough, a part of the population remains polarized. For our grid model, the sufficient probability for completely removing polarization from a fully polarized population seems to be between $h = 0.72$ and $h = 0.74$.



Extended Data Figure 6: The effect of global authorities on two polarized GANDHI factions is stable over time. a, The reputation difference (top tiles) and the average player score (bottom tiles) after 250 generations for several values of the probability h of encountering `virtue` and `evil` in a duel of two competing Gandhi factions. **b**, The number of polarized players that remain after 125 and 250 generations for varying values of h. Below the depolarization threshold of roughly 0.735, some polarized players remain present and continuously act as seed for new polarization; this fraction remains stable over time.



Extended Data Figure 7: The effect of local reciprocity on two competing factions of GANDHI over time. Strategies are randomly assigned at the start. Both factions follow the GANDHI strategy, but cooperate with any player that cooperated with them during the last encounter of these two players. **a**, Average score of players over time. **b**, Fraction of players seen as bad by both GANDHI factions over time. **c**, Reputation difference (top tiles) and last score (lower tiles) at different times of the simulation. Frequent strategy changes lead to some players becoming bad — they only cooperate with their neighbors due to direct reciprocity and hence cannot defect against bad players. However, direct reciprocity ensures that eventually, all players cooperate despite the bad reputation, which leads to a high average score.



Extended Data Figure 8: Gandhi++ loses to Gandhi in a direct competition. The strategy distribution (top tiles), reputation difference (middle tiles) and player score (bottom tiles) in a typical run of Gandhi++ (blue) against Gandhi (red) with probability $h = 0.1$ per duel for contact with the global authorities `Evil` and `Virtue`. The Gandhi++ population quickly collapses and is taken over by Gandhi.