Signals of Belonging: Emergence of Signalling Norms as Facilitators of Trust and Parochial Cooperation *

Ana Macanovic [a, b]
Milena Tsvetkova [c]
Wojtek Przepiorka [a]
Vincent Buskens [a, b]

[a] Department of Sociology / ICS, Utrecht University, Utrecht, The Netherlands

[b] Centre for Complex Systems Studies, Utrecht University, Utrecht, The Netherlands

[c] Department of Methodology, London School of Economics and Political Science, London, United Kingdom

[*] Corresponding authors: Ana Macanovic, a.macanovic@uu.nl; Wojtek Przepiorka, w.przepiorka@uu.nl

Acknowledgements: AM gratefully acknowledges financial support by the Swaantje Mondt travel fund.

Signals of Belonging: Emergence of Signalling Norms as Facilitators of Trust and Parochial Cooperation

Mechanisms of social control reinforce norms that appear harmful or wasteful, such as mutilation practices or extensive body tattoos. We suggest such norms arise to serve as signals that distinguish between ingroup "friends" and outgroup "foes", facilitating parochial cooperation. Combining insights from research on signalling and parochial cooperation, we incorporate a trust game with signalling in an agent-based model to study the dynamics of signalling norm emergence in groups with conflicting interests. Our results show that costly signalling norms emerge from random acts of signalling in minority groups that benefit most from parochial cooperation. Majority groups are less likely to develop costly signalling norms. Yet, norms that prescribe sending costless group identity signals can easily emerge in groups of all sizes – albeit, at times, at the expense of minority group members. Further, the dynamics of signalling norm emergence differ across signal costs, relative group sizes, and levels of ingroup assortment. Our findings provide theoretical insights into norm evolution in contexts where groups develop identity markers in response to environmental challenges that put their interests at odds with the interests of other groups. Such contexts arise in zones of ethnic conflict or during contestations of existing power relations.

Keywords: social norms, trust, cooperation, signalling, inter-group conflict, group identity

1. Introduction

At times, mechanisms of social control reinforce norms that appear individually, or even collectively, costly or "wasteful" – such as mutilation practices [1] or extensive body tattoos [2]. We test the conjecture that such norms emerge as outcomes of signalling games [3–6] in contexts where groups could benefit from parochial cooperation – i.e., cooperation with members of one's own group. The resulting signalling norms prescribe behaviours that mark individuals' belonging to a certain group, thereby helping distinguish between ingroup "friends" with aligned interests and outgroup "foes" with opposing interests [4,5]. We build on insights from literatures on costly signalling and parochial cooperation to shed light on the emergence of social norms prescribing the displaying of signs of group belonging.

Research on signalling has shown that costly behaviours that signal individuals' cooperative intent can be part of an (evolutionary stable) Nash equilibrium and enable observers of these signals to distinguish between cooperators and defectors [3,6–11]. These signals are reliable if only cooperators can afford to send them, either because cooperators incur lower signalling costs or gain higher cooperation benefits compared to untrustworthy defectors [12]. While some have suggested that norms prescribing costly signalling can emerge from arbitrary behaviours introduced by a single individual [4], others have indicated that more substantial "shocks" are needed to shift a population from a non-signalling to a signalling equilibrium [3,8]. At the same time, literature on parochial cooperation has shown that cooperation conditional on group belonging emerges and persists in the presence of inter-group competition or conflict [13,14] within groups of relatively small sizes [15]. This strand of literature has mostly assumed that cooperation can be conditioned on apparent innate features (i.e., conspicuous physical characteristics) that double as markers of group identity that are hard to fake [15–19].

We bring together similar – albeit often differently conceptualized – concepts from these two literatures to understand the conditions that favour the emergence of signalling norms as facilitators of parochial cooperation [5,20]. Following the literature on parochial cooperation, we put the emphasis on intergroup conflict and the resulting need to identify ingroup members and cooperate conditional on group identity. In line with the literature on signalling, we operationalize indicators of group identity as symbolic – and often costly – markers that can, but do not have to, be displayed by individuals [21,22]. In particular, we build on the trust game with signalling previously introduced to derive hypotheses about the emergence of signalling norms [5] and incorporate it in an agent-based model. Our agent-based model allows us to study the dynamics of norm emergence and parochial cooperation that intergroup conflict may bring about.

Our research question is: Under what conditions do signalling norms emerge to facilitate parochial cooperation?

Imagine two large hunter-gatherer bands that are in conflict over limited resources in their shared habitat [23]. A group that has spent a very long time hunting away from their band is coming home and approaching one of the settlements. Others in this band might want to welcome the hunters, hoping for a share in the kill. Yet, as the group has been away for so long, the band cannot as easily recognize which band this particular group belongs to. If they are from the same band, the incoming hunters will be cooperative; if they are from another band, the group might have come to raid the settlement instead. To prevent being mistaken for members of an enemy group and win the trust of their band, the approaching group of hunters can display a reliable sign of band belonging, such as a difficult to fake local dialect or a tattoo with group-specific features [24]. Once the band members recognize that this signal corresponds to their own group, they can rest assured and welcome the arrivals; otherwise, they can play safe and stay away from interacting with the group of strangers that is approaching their settlement.

In particular, we test the hypothesis that, in contexts where an outgroup is frequently encountered – such as in zones of ethnic conflict [20,25,26], prisons with rival gangs [27,28] or during contestations of existing power relations [29,30] – group members will be more willing to invest considerable resources in costly displays of group belonging [3,5,20,31,32]. We call behaviours that change or reinforce observers' beliefs about someone's belonging to a social group signalling norms [5]. We consider signals that are hard to fake (for instance, because they require inaccessible group-specific knowledge) and cannot be discriminated against by the outgroup (for example, because they are unknown to them [33,34]). We first provide analytical results on the conditions under which signal display and recognition can be preferred strategies of group members. We then employ agent-based simulations to understand how likely random changes in individual behaviours are to push populations from a non-signalling state into a state where the display of signals and their recognition are widely adopted – and the signalling norm is accepted [4]. Finally, we relax several model assumptions to understand an even broader range of conditions that favour the emergence of signalling norms.

Our study contributes to the literature on norm emergence and change in multiple ways. First, our agent-based model enhances our understanding of how signalling norms can solve trust dilemmas often arising in real-world exchange situations [12]. Second, we show that norms prescribing costly signalling of group identity can emerge and facilitate parochial cooperation in minority groups from initial conditions without signalling and signal recognition. Third, we show how costless signals of group identity support parochial cooperation in groups of all sizes. We thereby reveal a crucial difference in the dynamics of signalling trait emergence between norms of sending costly and costless signals. Fourth, we discuss how signalling norm emergence becomes more challenging for minority groups if signals can be recognized by the outgroup or there is significant noise in perceptions of an individual's group identity. Our findings shed light on various real-life situations, such as how dialects [35] or specific types of scars [22] can be used to distinguish between ethnic groups in contexts of ethnic conflict or resource competition. Further work in this domain can help understand how groups coordinate on specific signalling norms in similar contexts, and how established signalling norms can hamper cooperation between groups once intergroup conflict is resolved.

2. Theory

2.1. Modelling the trust game with two groups

We incorporate the trust game with signalling [5] in an agent-based model to map the boundary conditions for the emergence of signalling norms and their impact on individual and group outcomes. The trust game (Figure S1) is a sequential game where one individual (the truster) chooses whether or not to trust the other individual (the trustee), who then decides whether to honour or abuse that trust [36]. In the trust game with signalling, the trustee can signal their identity to the truster before the truster makes a decision on whether to trust the trustee.[1] Using a trust game to study the emergence of signalling has two main advantages over using other games such as the prisoner's dilemma. First, the trust game models sequential decisions which are characteristic of many real-word interactions (e.g., market exchanges or hiring decisions). Second, in the trust game, player roles are not interchangeable and payoffs are, therefore, asymmetric. Both the sequential nature and asymmetry of the trust game allow us to attach behaviour to either trust or cooperation; which is unlike simultaneous-move and symmetric dilemmas (such as prisoner's dilemma), where it is unclear whether it is fear or greed that drives behaviour [39].

We adopt the trust game with signalling [5] to model exchange relations between members of two groups with conflicting interests (Figure S2). We set the payoffs such that a trustee always has an incentive to honour the trust of an ingroup truster and abuse the trust of an outgroup truster. Hence, trusting an ingroup trustee results in a payoff of R_I for both parties [25], whereas trusting an outgroup trustee results in a payoff of S for the truster and T for the trustee. If trust is not given, there is no exchange and both parties receive payoff P. The payoffs are ordered as follows: $R_I > T > R_O > P > S$. R_O denotes the payoff of both parties if a trustee honours the trust of an outgroup truster. By setting $R_I > T > R_O$, we hard-code the conflicting group interests into our

[1] Trust game with signalling resembles the hostage trust game [37–39].

model, so that trustees do not have an incentive to honour outgroup trusters' trust. This allows us to examine the role of signalling in contexts where, for instance, environmental constraints make capturing resources from neighbouring groups attractive [23,28]. In our model, trustees can pay a cost c to signal their group identity and trusters can condition their trust on these signals.

Our model makes three important assumptions: (1) signals are hard or impossible to fake [31]; (2) recognizing signals requires group-specific knowledge, so that members of the outgroup do not recognize them [33]; (3) trusters reveal their group identity when placing trust, which allows trustees to (not) honour the trust conditional on truster's group belonging.[2] With these assumptions our model captures situations in which trusters, but not trustees, suffer from lack of information about the other side's group belonging (and, thereby, intentions). In Section 4 we discuss the consequences of relaxing these three assumptions.

2.2. Conditions for signalling norm existence

In this section, we outline the conditions under which trusters consider group identity signals and trustees bear the costs of sending them. Individuals are first assigned a role of a truster or a trustee at random and then randomly matched with another individual of the opposite role. An individual will meet a member of their ingroup with a probability α that is equal to the share of this group in the total population (p_i where $i \in \{1, 2\}$ refers to either of two groups). We build on the suggestion that, the lower the probability of encountering ingroup members, the more likely a group is to develop a signalling norm [5]. More precisely, with $\alpha \leq \alpha^*$ (in our setup: minority groups) trusters encounter their ingroup so rarely that they are best off distrusting any trustee in the absence of signalling. With $\alpha > \alpha^*$ (majority groups), trusters encounter their ingroup frequently enough to prefer trusting any trustee in the absence of signalling (Eq. 1 and 2 in Supplementary Material). In Theorem 1 below, we establish the conditions necessary for the equilibrium where, within a group, all group members signal and conditionally trust (i.e., give trust conditional on having observed the ingroup's signal). In Theorem 2, we establish the existence of an equilibrium without signalling and signal recognition. We provide proofs and the full analysis in Section S1 of the Supplementary Material. Finally, we define signalling norm emergence as the process during which a group in a non-signalling equilibrium moves to a signalling equilibrium.

[2] This could be because trustees acquire additional information about trusters after their signalling decision is made (e.g., the group from our example in the Introduction recognizes, by its appearance, the village they are approaching only after having decided to display a signal of their group identity) or because the truster reveals their group membership by the act of giving trust (e.g., by using a specific dialect when addressing the trustee).

Theorem 1. There exists a Nash Equilibrium where trusters condition their trusting on signals of the ingroup and trustees display group identity signals (signalling equilibrium) if and only if the signalling costs are offset by the benefits of parochial cooperation established with the help of signalling ($c \leq \alpha[R_I - P]$, where $R_I - P$ captures the benefits of cooperation).

Theorem 2. In a group with $\alpha \leq \alpha^*$, there exists a Nash Equilibrium where trusters unconditionally distrust all trustees and trustees do not display group identity signals (non-signalling equilibrium). In a group with $\alpha > \alpha^*$, there exists a non-signalling equilibrium where trusters unconditionally trust all trustees and trustees do not display group identity signals.

To understand how signalling norms emerge, we evaluate how likely a group is to move from a non-signalling into a signalling equilibrium from random changes in strategies of trustees or trusters (e.g., a trustee suddenly starts sending some signal of group identity). Note that we only consider a signalling norm to have emerged if trustees signal their identity *and* trusters recognize and condition their trust on these signals. We provide the full analysis in Section S1.2 of the Supplementary Material. Our analysis shows that trusters in a group where $\alpha \leq \alpha^*$ are willing to recognize signals, rather than unconditionally distrust all trustees, as long as the share β of signalling ingroup trustees satisfies the following condition:

$$\beta \geq \beta^* = 0 \tag{1}$$

In the group where $\alpha > \alpha^*$, trusters recognize signals, rather than unconditionally giving their trust to all trustees, if the share β of signalling ingroup trustees satisfies the following condition:

$$\beta \geq \beta^* = 1 - \frac{(1-\alpha)(P-S)}{\alpha(R_I - P)} \tag{2}$$

In both groups, trustees will be willing to bear the cost c of sending a signal of their group identity if the share γ of ingroup trusters who recognize ingroup signals satisfies the following condition:

$$\gamma \geq \gamma^* = \frac{c}{\alpha(R_I - P)} \tag{3}$$

These conditions show that the threshold β^* is always lower in groups with $\alpha \leq \alpha^*$. That is, trusters in minority groups are always willing to trust conditional on having observed the signal, which further ensures that the threshold γ^* needed for trustees to signal is more easily reached. Trusters in a group with $\alpha > \alpha^*$ only start conditioning their trusting behaviour on group identity signals if a sufficiently high number of trustees already signals their identity. We therefore expect that random changes in strategies of trustees and trusters will be more likely to shift the minority, rather than majority, groups into the signalling equilibrium.

2.3. Modelling signalling norm emergence in populations

Our formal analysis allows us to formulate expectations regarding the likelihood of groups to shift to the signalling equilibrium under certain conditions. However, to understand the dynamics of signalling norm emergence from random variations in individual agent behaviours across different conditions, we use an agent based model with social learning dynamics.

Agents in our model have two main traits that determine their strategy in each of the roles they can assume in the game: the signalling and the trusting trait. The signalling trait determines whether an agent signals their group identity at a predefined cost *c* when acting as a trustee. We vary signalling costs to capture a range of different behaviours that can convey information about group membership. The trusting trait defines whether an agent will, in the role of a truster, trust unconditionally, distrust unconditionally, or trust conditionally (upon having recognized a signal of their ingroup). There are six possible strategies resulting from different combinations of these two traits.

We simulate a population of 500 agents randomly assigned to group 1 with probability p_1 and group 2 with probability $p_2 = (1 - p_1)$. Group sizes and memberships are held fixed. Agents assume the role of a truster or a trustee with equal probability in each round and are then matched with another agent of the opposite role with whom they play the trust game once (if without a match, an agent sits out the round). As in our theoretical model, the probability of meeting one's ingroup α is equal to the share of ingroup p in the population. Additionally, we introduce parameter φ capturing assortative matching with regard to group membership: the probability of meeting the ingroup, holding p fixed, increases with φ. In our agent-based model α, thus, depends on both p and φ.[3] Agents cannot select whom they are matched with and have no memory of past encounters. Such a setup helps us model scenarios where group members frequently interact with strangers, which is a rather adverse environment for the evolution of parochial cooperation [40].

We set the trust game payoffs such that $\alpha^* = 0.43$ (see Figure 1 for payoffs and Supplementary Material for α^* calculation). Given this, we choose values of p to test scenarios with α both below and above this threshold. We do this so that the group with $\alpha \leq \alpha^*$ always constitutes a minority in the population and faces a group with $\alpha > \alpha^*$ that constitutes the majority in the population. As our focus is on the emergence of the two traits that constitute the signalling norm (signal sending and recognition), we initialize the model without any signalling ($\beta_1 = \beta_2 = 0$) or conditional trusting (i.e., signal recognition; $\gamma_1 = \gamma_2 = 0$) in either of the groups and randomly assign agents to

[3] In Table S2 in the Supplementary Material we show α values resulting from different combinations of p and φ parameters tested in our models.

unconditional trusting and unconditional distrusting in both groups ($\delta_1 \approx \delta_2 \approx 0.5$ and $\varepsilon_1 \approx \varepsilon_2 \approx 0.5$). We introduce random mutations of agents' traits with probability m_o, and vary this value to capture different propensities of agents to introduce new behaviours with regard to identity signalling [41].

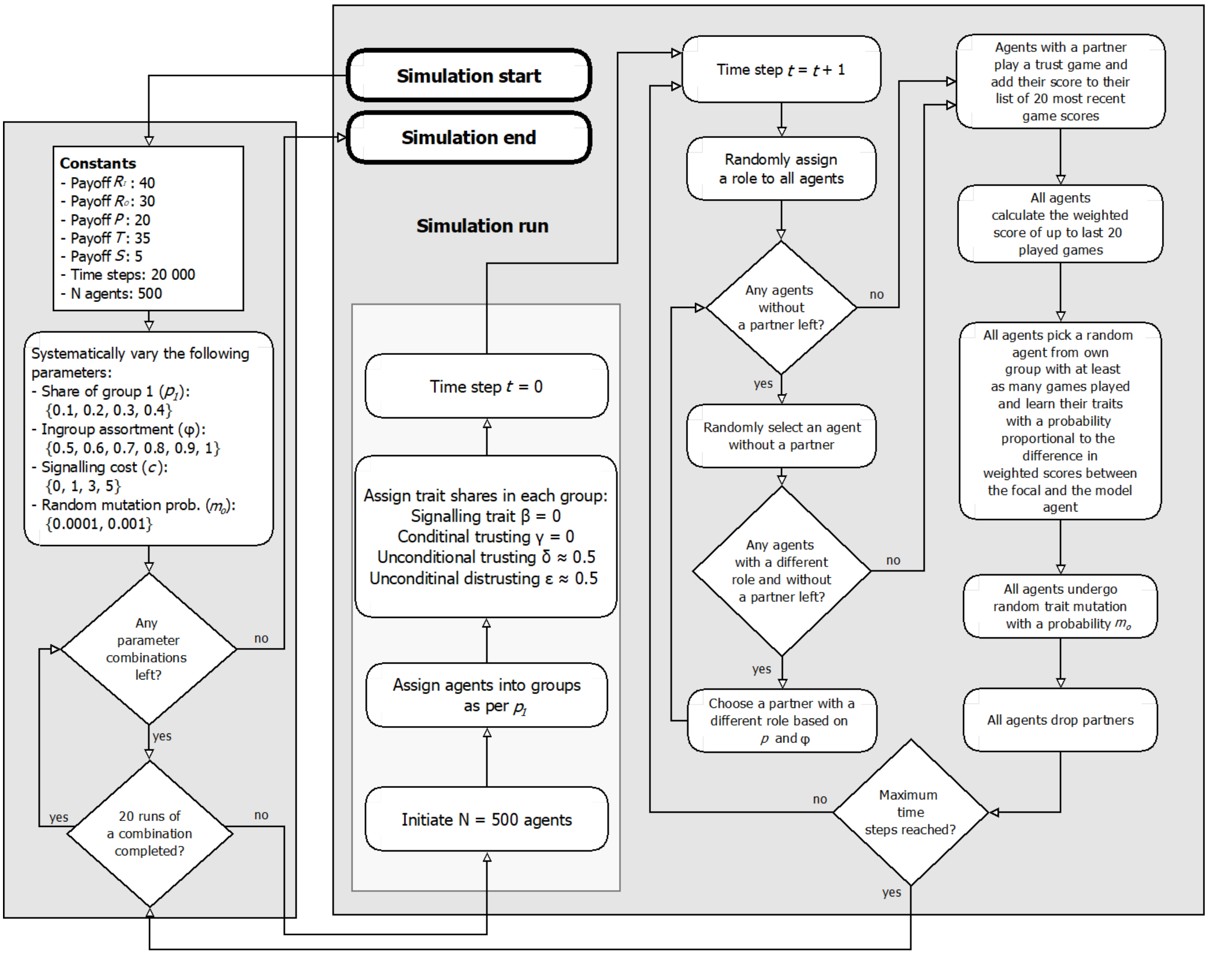

Simulation start
Simulation end
Constants
- Payoff R_I : 40
- Payoff R_O: 30
- Payoff P: 20
- Payoff T: 35
- Payoff S: 5
- Time steps: 20 000
- N agents: 500
Systematically vary the following parameters:
- Share of group 1 (p_1): {0.1, 0.2, 0.3, 0.4}
- Ingroup assortment (φ): {0.5, 0.6, 0.7, 0.8, 0.9, 1}
- Signalling cost (c): {0, 1, 3, 5}
- Random mutation prob. (m_o): {0.0001, 0.001}
Any parameter combinations left?
no
yes
20 runs of a combination completed?
yes
no
Simulation run
Time step t = 0
Assign trait shares in each group: Signalling trait β = 0 Conditinal trusting γ = 0 Unconditional trusting δ ≈ 0.5 Unconditinal distrusting ε ≈ 0.5
Assign agents into groups as per p_1
Initiate N = 500 agents
Time step $t = t + 1$
Randomly assign a role to all agents
Any agents without a partner left?
no
yes
Randomly select an agent without a partner
Any agents with a different role and without a partner left?
no
yes
Choose a partner with a different role based on p and φ
Agents with a partner play a trust game and add their score to their list of 20 most recent game scores
All agents calculate the weighted score of up to last 20 played games
All agents pick a random agent from own group with at least as many games played and learn their traits with a probability proportional to the difference in weighted scores between the focal and the model agent
All agents undergo random trait mutation with a probability m_o
All agents drop partners
Maximum time steps reached?
no
yes

Figure 1. Flowchart of the simulation process. For each combination of parameters, we run twenty independent randomly initiated simulations. After all available agents have played the trust game, all agents undergo social learning and random mutation before the simulation moves onto the next time step – for a maximum of 20 000 steps.

After playing a trust game, agents can update their strategies by observing and copying both of the traits of their more successful ingroup members in a mechanism resembling social learning [42,43][4] and, thereafter, can also undergo a random mutation to one of their traits. We simulate the evolution of agent traits for 20 000 time steps, evaluating each combination of parameters (see Table S3) in twenty independently initialized simulation runs. Figure 1 gives an overview of the full simulation procedure. Finally, we also run reference simulations for 2 500 time steps

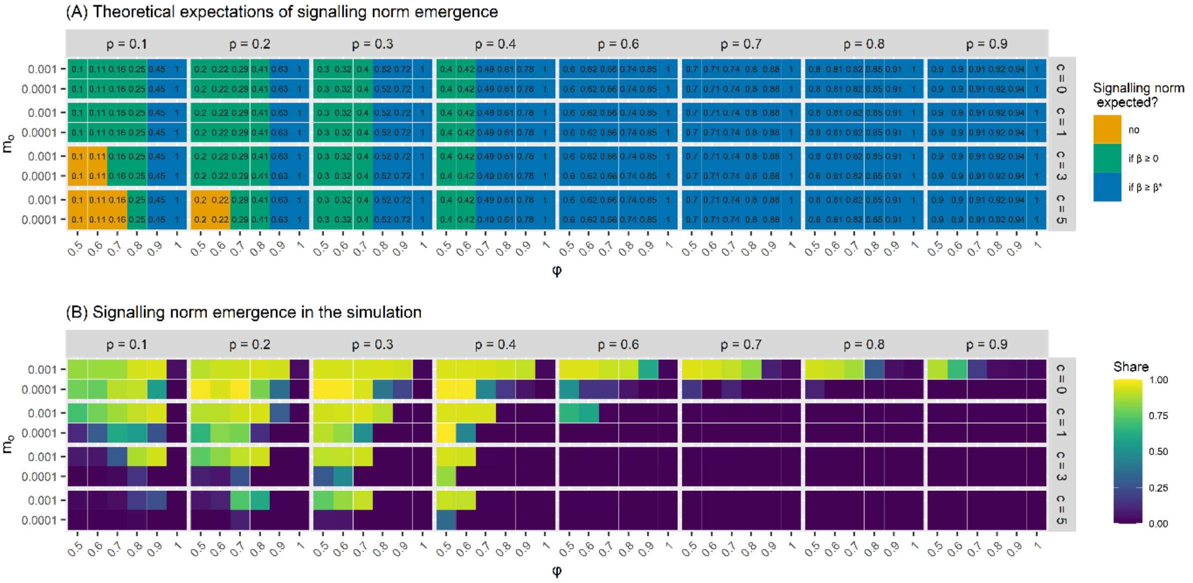
[4] Agents compare their weighted payoffs from the last twenty rounds to the equivalent payoffs of a randomly chosen ingroup agent in the same role who has played a comparable number of games; the more successful the model agent, the more likely the focal agent is to adopt their traits (see Section S3.1 in the Supplementary Material). We exclude learning from the outgroup since the two groups' interests are in conflict.

following the same procedure, but excluding a possibility of signalling norm emergence. These simulations serve as a baseline for comparing the outcomes of the signalling equilibrium to those of a non-signalling one (Figure S3).

3. Simulation results

3.1. Signalling and parochial cooperation

In Panel A of Figure 2, we show the expectations regarding signalling norm emergence according to our formal analysis (Eq. 1 and Eq. 2 and Section S1.2 of the Supplementary Material). In Panel B, we show the results of our agent-based simulations.[5] Panel B shows the prevalence of the signalling norm (i.e., agent strategy including both a signalling and a conditionally trusting trait) within each group averaged across the last 100 rounds of simulations. Henceforth, we refer to groups with $p < 0.5$ as minority groups, and groups with $p > 0.5$ as majority groups. For ease of interpretation, we focus on cases when $\varphi = 0.5$ (therefore, $\alpha_i = p_i$) unless stated otherwise.


(A) Theoretical expectations of signalling norm emergence

(B) Signalling norm emergence in the simulation

Figure 2. Panel A shows theoretical expectations regarding the signalling norm emergence given our analytical results. Tile colours correspond to different emergence conditions (orange corresponds to the condition in Theorem 1, green in Eq. 1 and blue in Eq. 2) and tile labels correspond to the α values resulting from different combinations of p and φ. Panel B shows the results of our agent-based simulations. Tile colouring reflects the share of agents with a strategy including both signalling and trusting conditional on signals averaged across the last 100 rounds given the share of ingroup (p) and the assortment parameter (φ) under varying signal costs c and random mutation parameters. We average the results over twenty independent runs for each parameter combination. See Figures S4 and S5 for an overview of shares of other strategies.

[5] In Section S4.9 of the Supplementary Material, we provide evidence of the robustness of our simulations to changes in several parameters.

As we predict based on our analytical results, minority groups develop signalling norms (as $\beta \geq 0$) when signal costs are low enough.[6] For example, when $c = 1$ and random mutations are high ($m_o = 0.001$) almost all minority groups fully develop a signalling norm (second row in Panel B, upper tick on the Y axis). In minority groups, trusters are indifferent between distrusting unconditionally and trusting upon observing the signal (Eq. 1). Thus, if a few trustees start signalling (following a random mutation), they are likely to encounter trusters who have started – and continued – conditionally trusting frequently enough to reap the benefits of parochial cooperation. These benefits offset the costs of signalling, allowing social learning to spread both signalling and conditional trusting in the group. However, when $c > 1$, groups with $\alpha \leq 0.2$ cannot afford to send signals.[7] Furthermore, if mutation rates are low ($m_o = 0.0001$), the signalling norm spreads more slowly and does not become as widely adopted (also see next sub-section).

In most cases, signalling norms do not emerge in majority groups with $\alpha > \alpha^*$ if signals are costly. Trusters in these groups are unconditionally trusting and only willing to condition their trusting behaviour on a signal if the share of signalling trustees is sufficiently high (Eq. 2). Therefore, even if some trustees do start signalling, social learning is likely to bring any trusters who start conditionally trusting back to unconditional trust instead; in turn, as long as the share of conditionally trusting trusters remains too low (Eq. 3), the number of trustees willing to bear the costs of sending signals will remain low as well.[8] Yet, when signals are costless ($c = 0$), trustees who start signalling will have no incentive to stop doing so and signalling can spread through social learning. This, in turn, makes conditional trusting more attractive for trusters and establishes the signalling norm in some majority groups as long as their α is sufficiently low (otherwise, a very large share of signalling is needed for conditional trust to proliferate) and the random mutations are sufficiently high (topmost row in Panel B of Figure 2).

Figure 2 also shows that smaller minority groups adopt signalling more, and larger minority groups less, at higher levels of assortment (i.e., as their α increases, see also Figure 2A and Table S2 for easier interpretation). We find that, overall, groups with similar α values stemming from different combinations of p and φ develop comparable shares of the signalling and conditional trusting strategy (Figure S6). Yet, as we show in the next subsection, dynamics of signalling norm emergence can differ depending on the group size.

Overall, groups that develop a (costly) signalling norm are better off than they were in the absence of signalling (see Panel A of Figure S7 for the average payoffs obtained by agents within each

[6] See Section S4.6 of the Supplementary Material for results from simulations initiated with full signalling and signal recognition in both groups. These results are in alignment with our theoretical expectations in Figure 2A.
[7] Recall from Theorem 1 that costs need to satisfy the condition $c \leq \alpha(R_I - P)$ to be affordable for trustees.
[8] For instance, for a group with $p = 0.8$, the threshold value as per Eq. 2 would be $\beta^* \geq 0.81$ (Table S4).

group compared to a non-signalling baseline scenario). Yet, while a group's signalling and conditionally trusting behaviours are not dependent on the outgroup's strategies in our model, changes in outgroup trait distributions still affect group payoffs. For instance, despite obtaining benefits from establishing parochial cooperation with the help of signals, minority groups net losses when signals are costless (c = 0). This is due to the fact that the emergence of signalling norms (and parochial cooperation) in majority groups closes the door for the minority groups' exploitation of the majority. When signals are costly, despite bearing the costs of sending signals, minority groups that adopt signalling and conditional trusting see an increase in average payoffs per agent that mainly stems from truster benefits (see Panels B and C in Figure S7). However, if signalling behaviour appears in the group, but is not followed by sufficient conditional trusting (e.g., red tiles when p = 0.6 and c > 0 in Panel A of Figure S7), groups net losses because trustees bear the costs of signalling, but do not reap the benefits of parochial cooperation.

3.2. Dynamics of signalling norm emergence

To understand the dynamics of norm emergence we show trait change dynamics in Figure 3, comparing a scenario with costless signals (Panel A) and a scenario with signals of intermediate cost (Panel B). For ease of interpretation, we only consider scenarios without any group assortment (φ = 0.5, α = p); Figures S9-S12 show detailed trait evolution plots with 95% confidence intervals. In Panel A in Figure 3, we show that, if signals are costless (c = 0), moderate minority groups (p ≥ 0.2) adopt the signalling trait the fastest, followed by moderate majority groups (0.6 ≤ p ≤ 0.7). Costless signalling fixates quickly, followed by conditional trusting; yet, the latter spreads with a smaller delay in minority than in majority groups. These results are in line with the results from our analytical model (Eq. 2): in groups with $\alpha > \alpha^*$, conditional trusting only pays off once a sufficiently high share of ingroup trustees signal their identity.

Panel B shows that, under intermediate signal costs (c = 3), most minority groups (0.2 ≥ p ≤ 0.4) develop signalling and conditional trusting traits. Majority groups do not adopt conditional trusting as the share of signalling trustees never reaches the threshold needed for conditional trusting to be a viable strategy (Eq. 2). Note that, unlike in the costless signalling scenario, the proliferation of signalling traits depends on (and lags behind) a sufficient supply of conditional trusting.

In Figure 2 we showed that signalling norms do not fixate when mutation rates are low (m_o = 0.0001). Inspecting the dynamics of these simulations suggests that signalling norms develop in some simulation runs, but not in others. Signalling takes off only once a sufficiently high share of trusters adopts signal recognition; but, depending on the dynamics of individual runs, this might not happen in every run (five out of twenty runs when p = 0.3, φ = 0.5, and c = 3, as shown in

Figure S14).[9] The lower propensity of agents to randomly adopt signalling and conditional trusting makes it more difficult for social learning to boost the coevolution of the two traits in a manner that supports the wide adoption of the signalling norm.

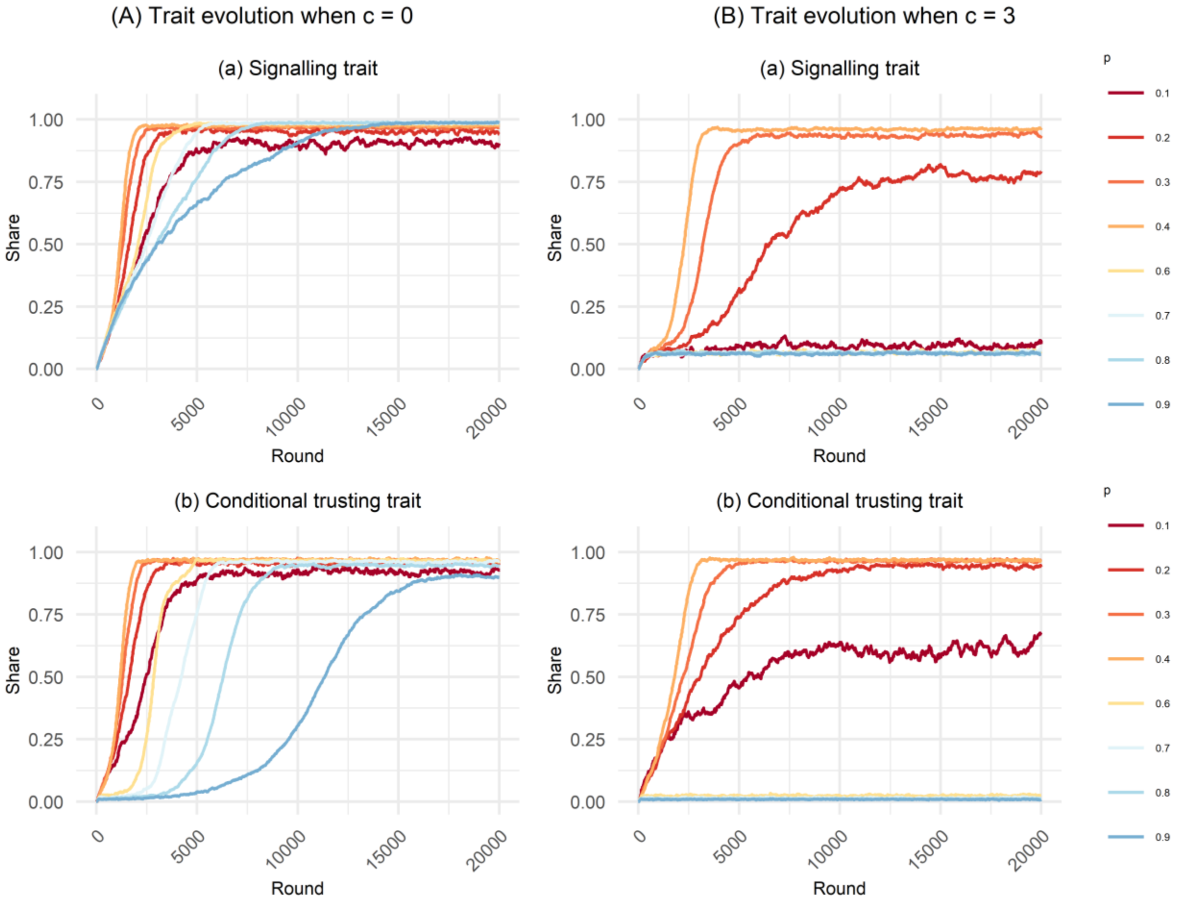


Figure 3. Evolution of signalling (a) and conditional trusting (b) traits over 20 000 rounds if signals are costless (Panel A, *c* = 0) or costly (Panel B, *c* = 3), there is no assortment (φ = 0.5) and random mutations are more common (m_o = 0.001). Shares on the Y axis represent share of different traits within each group. We average the results over twenty independent runs for each parameter combination.

Inspecting different combinations of p and φ that result in the same α value, we observe similar rates, but different dynamics of signalling norm emergence. Smaller minority groups with a higher ingroup assortment (e.g., p = 0.3 and φ = 0.7) take longer to develop the signalling norm and show more variability across runs compared to larger groups with lower assortment, and same α (e.g., p = 0.4 and φ = 0.5, α = 0.4 for both, see Figures S17 and S18).

[9] Figure S16 shows how, in a similar scenario with c = 5, a signalling norm emerges in only one out of 20 runs by time step 20 000.

4. Relaxing model assumptions

Here, we briefly discuss how relaxing some of the model assumptions affects the conditions under which signalling norms emerge. Full results are discussed at length in Section S2 of the Supplementary Material. We relax each of the assumptions individually, and not simultaneously. First, we relax the assumption that trustees can only send ingroup signals, allowing them to send outgroup signals in order to exploit the conditionally trusting outgroup. However, sending outgroup signals is not an equilibrium strategy for most groups, minority and majority alike. Only in the cases when a rather small minority group – whose benefits of parochial cooperation are low – faces a rather large majority group in a signalling equilibrium will there be an equilibrium with outgroup signalling. In these cases, minority trustees send outgroup signals (while minority trusters unconditionally distrust) and majority trusters conditionally trust (thereby reaping the benefits of parochially cooperating with the signalling ingroup and, at times, bearing the losses of being exploited by the outgroup).

Second, we relax the assumption that outgroup trusters do not recognize group signals, allowing trusters distrust only those trustees who display an outgroup signal. Signalling and signal recognition is still an equilibrium in several cases: (1) if both groups are in a signalling equilibrium; (2) if one group is in a signalling equilibrium, whereas trustees in the other do not signal and trusters discriminate against outgroup signals. In this setup, trustees are willing to bear lower costs of signalling compared to our main analysis. This is because signalling now includes a possibility of being distrusted by outgroup trusters – who could be abused without consequences in the main model. Our simulation results show that signalling norms still emerge in this setup, but in a somewhat narrower set of minority groups (Figures S20 and S21).

Finally, we relax the assumption that trustees always perceive the trusters' group identity correctly upon being given trust. We define the threshold of error (i.e., noise) in the accuracy of trustees' perception of truster identity below which signalling and signal recognition remain equilibrium strategies. Our analytical results suggest this is the case even when trustees' perception of truster group identity is as good as a random guess. However, very high levels of noise can hamper signalling norm emergence as relative group size decreases or signalling cost increases (Figure S22). The level of noise groups of different sizes can tolerate in a signalling equilibrium depends on the payoff structure.

5. Discussion and conclusion

We combine insights from literatures on signalling and parochial cooperation to understand how signalling norms evolve to facilitate parochial cooperation in groups with conflicting interests.

Using a trust game allows us to model encounters where giving trust to an untrustworthy other exposes one to significant risks, while not giving trust can prevent one from successfully cooperating with the ingroup. Our formal analysis suggests that, while following a signalling norm can be an equilibrium for different groups, groups who face a lower share of potential ingroup cooperators are more likely to shift from a non-signalling to a signalling equilibrium. We use agent-based simulations to examine the emergence of traits related to signalling norms under different signalling costs, mutation rates, relative group shares, and extent of inter-group assortment.

Our agent-based simulations confirm that groups that constitute the minority in the population – and are, thus, less likely to encounter cooperative ingroup members – more easily develop (costly) signalling norms. Signalling norms facilitate cooperation with the ingroup while avoiding exploitation by the outgroup. However, when ingroup assortment is higher – for instance, because members of small groups live close to each other – signalling becomes less beneficial for promoting parochial cooperation [44]. We show that the sufficiently wide recognition of signals (i.e., conditional trusting) opens the door for costly signalling.

Broadly in line with the literature on tag-based cooperation [45], we find that costless signals that cannot be recognized by the outgroup emerge to facilitate parochial cooperation rather easily across the groups. Costless signals improve majority group benefits as they allow the majority to avoid exploitation by the minority. This is, however, at the expense of minority groups who lose the opportunity to exploit otherwise trusting outgroup members. These results showcase how, in setups where groups have conflicting interests, group benefits sensitively depend on the actions of the outgroup (also see [46]).

These findings remain robust even when relaxing several of our model assumptions. We assume that sending a signal of group identity requires cultural knowledge not easily accessible to the outgroup, which makes them hard to “fake” [21]. Yet, even if we allow groups to send outgroup signals, this strategy can benefit only rather small minority groups who are facing a majority outgroup who has already developed a signalling norm. Thereby, we specify the conditions under which outsiders could benefit from learning the (costly) “secret handshakes” of a group and exploit parochial cooperators by defecting instead of cooperating [47,48]. In our setup, sufficiently large groups can sustain parochial cooperation in spite of such exploitation. Extensions of our model could further explore other scenarios, such as those where groups can adopt new signals and establish cooperation before the signal is “hijacked” again, resulting in cycles of dominance of different reliable identity markers [18,19,21].

Our model considers signals that are so group specific that they remain undetectable to the outgroup [33,34]. If signals are unknown to the outgroup, group members can both reap the benefits of parochial cooperation with the ingroup and exploit the unsuspecting outgroup. Relaxing this assumption, we find that, if signals are easily recognized and the outgroup can discriminate against them, smaller minority groups become less likely to develop signalling norms as they cannot bear the losses from discrimination. Small minority groups will likely benefit from developing signals that are not as easily recognized by the outgroup. In addition, our simulations show that the emergence of signalling norms in most groups remain undisturbed even as we introduce significant amounts of noise to the perception of group identities during encounters. It is the smallest minority groups that are the most sensitive to increases in uncertainty about the counterpart's identity.

Existing work has suggested that incidental individual behaviours can develop into social norms prescribing signals of some qualities in the population [4]. Our results show that this can indeed be the case in groups that constitute a minority in the population as long as these incidental behaviours are frequent enough. Further, majority groups in the population cannot move from the non-signalling equilibrium. This is because they fail to reach the high threshold as from which sending costly signals becomes beneficial. Developing a signalling norm in majority groups might, thus, require a stronger stochastic shock to introduce sufficiently wide signal recognition, for instance, via a centralized or persuasive intervention [8,49,50].

Research has suggested that selectively interacting with one's group allows parochial cooperation to emerge [44,51–54]. We, however, allow agents to condition cooperation on the partner's signal (or absence thereof) in a randomly mixed population. This allows us to model how signals can emerge to support parochial cooperation even in adverse contexts where individuals have to face unknown others – as is often the case in large, complex societies where one cannot rely on knowledge from past interactions [40]. We further explore the interaction between relative group size and the likelihood of interacting with the ingroup. Our results highlight that isolated minority communities who infrequently encounter the outgroup benefit less from signals of group identity. At the same time, minority groups situated at group boundaries who encounter the outgroup more frequently obtain more benefit from being able to tell friends and foes apart. Even holding the probability of encountering the ingroup constant, smaller minority groups take longer to establish signalling norms compared to larger minority groups.

We analyse when a norm prescribing an individual signal of group identity can emerge to facilitate parochial cooperation. Combining our model with existing work that evaluates how groups coordinate on a specific signal among the multitude of potential candidates could help understand signalling norm emergence more generally [4,55–59]. To gain a more realistic understanding of how large numbers of group members come to recognize a specific signal as a reliable marker of group identity, studies could allow agents to choose their partners based on displayed signals or consider the possibility that certain individuals can send signals at a lower cost, kick-starting their recognition in the population [32,55].

Further, our model assumes that coherent groups with conflicting interests exist before signalling norms emerge. This is the case when, for instance, geographical boundaries or kinship ties determine group formation and constrain resource sharing with other groups. Generalizing our insights even further, future work could consider modelling the coevolution of groups and signalling norms [60]. Finally, our model captures contexts where individuals successfully cooperate within groups, but face environmental restrictions that put their interests at odds with those of other groups [23,30,61]. Extending our model to situations where periods of intergroup conflict are interrupted by peaceful coexistence can help understand how, by supporting parochial cooperation, signalling norms that emerge in the times of conflict could hamper intergroup cooperation during peaceful times [23].

6. References

1. Mackie G. 1996 Ending Footbinding and Infibulation: A Convention Account. *Am. Sociol. Rev.* **61**, 999. (doi:10.2307/2096305)

2. Gambetta D. 2009 *Codes of the Underworld: How Criminals Communicate*. Princeton, NJ: Princeton University Press.

3. Henrich J. 2009 The evolution of costly displays, cooperation and religion. *Evol. Hum. Behav.* **30**, 244–260. (doi:10.1016/j.evolhumbehav.2009.03.005)

4. Posner EA. 1998 Symbols, Signals, and Social Norms in Politics and the Law. *J. Leg. Stud.* **27**, 765–797. (doi:10.1086/468042)

5. Przepiorka W, Diekmann A. 2021 Parochial cooperation and the emergence of signalling norms. *Philos. Trans. R. Soc. B Biol. Sci.* **376**. (doi:10.1098/rstb.2020.0294)

6. Smith EA, Bliege Bird R. 2005 Costly Signaling and Cooperative Behavior. In *Moral Sentiments and Material Interests: The Foundations of Cooperation in Economic Life* (eds H Gintis, S Bowles, R Boyd, E Fehr), pp. 115–148. Cambridge, MA: MIT Press.

7. Barclay P, Bliege Bird R, Roberts G, Számadó S. 2021 Cooperating to show that you care: costly helping as an honest signal of fitness interdependence. *Philos. Trans. R. Soc. B Biol. Sci.* **376**, 20200292. (doi:10.1098/rstb.2020.0292)

8. Gintis H, Smith EA, Bowles S. 2001 Costly Signaling and Cooperation. *J. Theor. Biol.* **213**, 103–119. (doi:10.1006/jtbi.2001.2406)

9. Harms W, Skyrms B. 2009 Evolution of Moral Norms. In *The Oxford Handbook of Philosophy of Biology* (ed M Ruse), pp. 434–450. Oxford University Press. (doi:10.1093/oxfordhb/9780195182057.003.0019)

10. Smith EA, Bliege Bird RL. 2000 Turtle hunting and tombstone opening. *Evol. Hum. Behav.* **21**, 245–261. (doi:10.1016/S1090-5138(00)00031-3)

11. Fehrler S, Przepiorka W. 2013 Charitable giving as a signal of trustworthiness: Disentangling the signaling benefits of altruistic acts. *Evol. Hum. Behav.* **34**, 139–145. (doi:10.1016/j.evolhumbehav.2012.11.005)

12. Przepiorka W, Berger J. 2017 Signaling Theory Evolving: Signals and Signs of Trustworthiness in Social Exchange. In *Social dilemmas, institutions, and the evolution of cooperation* (eds B Jann, W Przepiorka), Berlin, Germany and Boston, MA: De Gruyter. (doi:10.1515/9783110472974-018)

13. Choi J-K, Bowles S. 2007 The Coevolution of Parochial Altruism and War. *Science* **318**, 636–640. (doi:10.1126/science.1144237)

14. De Dreu CKW, Balliet D, Halevy N. 2014 Parochial Cooperation in Humans: Forms and Functions of Self-Sacrifice in Intergroup Conflict. In *Advances in Motivation Science*, pp. 1–47. Elsevier. (doi:10.1016/bs.adms.2014.08.001)

15. Antal T, Ohtsuki H, Wakeley J, Taylor PD, Nowak MA. 2009 Evolution of cooperation by phenotypic similarity. *Proc. Natl. Acad. Sci.* **106**, 8597–8600. (doi:10.1073/pnas.0902528106)

16. Hammond RA, Axelrod R. 2006 The Evolution of Ethnocentrism. *J. Confl. Resolut.* **50**, 926–936. (doi:10.1177/0022002706293470)

17. Hartshorn M, Kaznatcheev A, Shultz T. 2013 The Evolutionary Dominance of Ethnocentric Cooperation. *J. Artif. Soc. Soc. Simul.* **16**, 7. (doi:10.18564/jasss.2176)

18. Jansen VAA, van Baalen M. 2006 Altruism through beard chromodynamics. *Nature* **440**, 663–666. (doi:10.1038/nature04387)

19. Riolo RL, Cohen MD, Axelrod R. 2001 Evolution of cooperation without reciprocity. *Nature* **414**, 441–443. (doi:10.1038/35106555)

20. Blanton RE. 2015 Theories of ethnicity and the dynamics of ethnic change in multiethnic societies. *Proc. Natl. Acad. Sci.* **112**, 9176–9181. (doi:10.1073/pnas.1421406112)

21. Cohen E. 2012 The Evolution of Tag-Based Cooperation in Humans: The Case for Accent. *Curr. Anthropol.* **53**, 588–616. (doi:10.1086/667654)

22. Sosis R, Kress H, Boster J. 2007 Scars for war: evaluating alternative signaling explanations for cross-cultural variance in ritual costs. *Evol. Hum. Behav.* **28**, 234–247. (doi:10.1016/j.evolhumbehav.2007.02.007)

23. De Dreu CKW, Gross J, Fariña A, Ma Y. 2020 Group Cooperation, Carrying-Capacity Stress, and Intergroup Conflict. *Trends Cogn. Sci.* **24**, 760–776. (doi:10.1016/j.tics.2020.06.005)

24. Moffett MW. 2013 Human Identity and the Evolution of Societies. *Hum. Nat.* **24**, 219–267. (doi:10.1007/s12110-013-9170-3)

25. Castro L, Toro MA. 2007 Mutual benefit cooperation and ethnic cultural diversity. *Theor. Popul. Biol.* **71**, 392–399. (doi:10.1016/j.tpb.2006.10.003)

26. McElreath R, Boyd R, Richerson PJ. 2003 Shared Norms and the Evolution of Ethnic Markers. *Curr. Anthropol.* **44**, 122–130. (doi:10.1086/345689)

27. Demello M. 1993 The Convict Body: Tattooing Among Male American Prisoners. *Anthropol. Today* **9**, 10. (doi:10.2307/2783218)

28. Doğan G, Glowacki L, Rusch H. 2022 Are strangers just enemies you have not yet met? Group homogeneity, not intergroup relations, shapes ingroup bias in three natural groups. *Philos. Trans. R. Soc. B Biol. Sci.* **377**, 20210419. (doi:10.1098/rstb.2021.0419)

29. Mann M. 1993 *The Sources of Social Power*. 1st edn. Cambridge University Press. (doi:10.1017/CBO9780511570902)

30. Kroneberg C, Wimmer A. 2012 Struggling over the Boundaries of Belonging: A Formal Model of Nation Building, Ethnic Closure, and Populism. *Am. J. Sociol.* **118**, 176–230. (doi:10.1086/666671)

31. Cohen E, Haun D. 2013 The development of tag-based cooperation via a socially acquired trait. *Evol. Hum. Behav.* **34**, 230–235. (doi:10.1016/j.evolhumbehav.2013.02.001)

32. Dumas M, Barker JL, Power EA. 2021 When does reputation lie? Dynamic feedbacks between costly signals, social capital and social prominence. *Philos. Trans. R. Soc. B Biol. Sci.* **376**, 20200298. (doi:10.1098/rstb.2020.0298)

33. Smaldino PE, Flamson TJ, McElreath R. 2018 The Evolution of Covert Signaling. *Sci. Rep.* **8**, 4905. (doi:10.1038/s41598-018-22926-1)

34. van der Does T, Galesic M, Dunivin ZO, Smaldino PE. 2022 Strategic identity signaling in heterogeneous networks. *Proc. Natl. Acad. Sci.* **119**. (doi:10.1073/pnas.2117898119)

35. Bugarski R. 2012 Language, identity and borders in the former Serbo-Croatian area. *J. Multiling. Multicult. Dev.* **33**, 219–235. (doi:10.1080/01434632.2012.663376)

36. Dasgupta P. 1988 Trust as a Commodity. In *Trust: Making and Breaking Cooperative Relations* (ed D Gambetta), pp. 49–72. Oxford, UK: Blackwell Publishing.

37. Raub W. 2004 Hostage Posting as a Mechanism of Trust: Binding, Compensation, and Signaling. *Ration. Soc.* **16**, 319–365. (doi:10.1177/1043463104044682)

38. Raub W, Keren G. 1993 Hostages as a commitment device. A game-theoretic model and an empirical test of some scenarios. *J. Econ. Behav. Organ.* **21**, 43–67. (doi:10.1016/0167-2681(93)90039-R)

39. Raub W, Weesie J. 2000 Cooperation via Hostages. *Anal. Krit.* **22**, 19–43. (doi:10.1515/auk-2000-0102)

40. Smaldino PE. 2019 Social identity and cooperation in cultural evolution. *Behav. Processes* **161**, 108–116. (doi:10.1016/j.beproc.2017.11.015)

41. Bliege Bird R, Smith EA. 2005 Signaling Theory, Strategic Interaction, and Symbolic Capital. *Curr. Anthropol.* **46**, 221–248. (doi:10.1086/427115)

42. Masuda N, Ohtsuki H. 2007 Tag-based indirect reciprocity by incomplete social information. *Proc. R. Soc. B Biol. Sci.* **274**, 689–695. (doi:10.1098/rspb.2006.3759)

43. Rendell L *et al.* 2010 Why Copy Others? Insights from the Social Learning Strategies Tournament. *Science* **328**, 208–213. (doi:10.1126/science.1184719)

44. García J, van den Bergh JCJM. 2011 Evolution of parochial altruism by multilevel selection. *Evol. Hum. Behav.* **32**, 277–287. (doi:10.1016/j.evolhumbehav.2010.07.007)

45. García J, van Veelen M, Traulsen A. 2014 Evil green beards: Tag recognition can also be used to withhold cooperation in structured populations. *J. Theor. Biol.* **360**, 181–186. (doi:10.1016/j.jtbi.2014.07.002)

46. Helbing D, Johansson A. 2010 Cooperation, Norms, and Revolutions: A Unified Game-Theoretical Approach. *PLoS ONE* **5**, e12530. (doi:10.1371/journal.pone.0012530)

47. Miller JH, Butts CT, Rode D. 2002 Communication and cooperation. *J. Econ. Behav. Organ.* **47**, 179–195. (doi:10.1016/S0167-2681(01)00159-7)

48. Nettle D. 1997 Social Markers and the Evolution of Reciprocal Exchange. *Curr. Anthropol.* **38**, 93–99. (doi:10.1086/204588)

49. Gavrilets S. 2020 The dynamics of injunctive social norms. *Evol. Hum. Sci.* **2**, e60. (doi:10.1017/ehs.2020.58)

50. Gavrilets S, Richerson PJ. 2022 Authority matters: propaganda and the coevolution of behaviour and attitudes. *Evol. Hum. Sci.* **4**, e51. (doi:10.1017/ehs.2022.48)

51. Fletcher JA, Doebeli M. 2009 A simple and general explanation for the evolution of altruism. *Proc. R. Soc. B Biol. Sci.* **276**, 13–19. (doi:10.1098/rspb.2008.0829)

52. Kim J-W, Hanneman RA. 2014 Coevolutionary Dynamics of Cultural Markers, Parochial Cooperation, and Networks. *J. Confl. Resolut.* **58**, 226–253. (doi:10.1177/0022002712468691)

53. Riolo RL. 1997 The Effects of Tag-Mediated Selection of Partners in Evolving Populations Playing the Iterated Prisoner's Dilemma. *SFI Work. Pap. 97-02-016 St. Fe NM*

54. Takesue H. 2020 From defection to ingroup favoritism to cooperation: simulation analysis of the social dilemma in dynamic networks. *J. Comput. Soc. Sci.* **3**, 189–207. (doi:10.1007/s42001-019-00058-4)

55. Barker JL, Power EA, Heap S, Puurtinen M, Sosis R. 2019 Content, cost, and context: A framework for understanding human signaling systems. *Evol. Anthropol. Issues News Rev.* **28**, 86–99. (doi:10.1002/evan.21768)

56. Bell AV. 2020 A measure of social coordination and group signaling in the wild. *Evol. Hum. Sci.* **2**, e34. (doi:10.1017/ehs.2020.24)

57. Centola D, Baronchelli A. 2015 The spontaneous emergence of conventions: An experimental study of cultural evolution. *Proc. Natl. Acad. Sci.* **112**, 1989–1994. (doi:10.1073/pnas.1418838112)

58. Przepiorka W, Szekely A, Andrighetto G, Diekmann A, Tummolini L. 2022 How Norms Emerge from Conventions (and Change). *Socius Sociol. Res. Dyn. World* **8**, 237802312211245. (doi:10.1177/23780231221124556)

59. Schelling TC. 1980 *The Strategy of Conflict*. Cambridge, MA and London, England: Harvard University.

60. Efferson C, Lalive R, Fehr E. 2008 The Coevolution of Cultural Groups and Ingroup Favoritism. *Science* **321**, 1844–1849. (doi:10.1126/science.1155805)

61. Rodrigues AMM, Barker JL, Robinson EJH. 2022 From inter-group conflict to inter-group cooperation: insights from social insects. *Philos. Trans. R. Soc. B Biol. Sci.* **377**, 20210466. (doi:10.1098/rstb.2021.0466)

Supplementary Material:

Signals of Belonging: Emergence of Signalling Norms as Facilitators of Trust and Parochial Cooperation

Supplementary Material for:

Macanovic, A., Tsvetkova, M., Przepiorka, W., and Buskens, V. 2023. Signals of Belonging: Emergence of Signalling Norms as Facilitators of Trust and Parochial Cooperation. *Philosophical Transactions of the Royal Society B.* (forthcoming), doi: 10.1098/rstb.2023.0029

S1. Game-theoretic model

We first describe our setup with a trust game with incomplete information and without signalling (Figure S1). We assume the existence of two groups with conflicting interests in the population. In this game, nature (denoted N in Figure S1) first determines the type of the trustee that a truster encounters as either ingroup (with the probability α) or outgroup (with the probability $1 - \alpha$).[1] The truster first decides whether they want to trust the trustee. If they decide to trust the trustee, the trustee then decides whether to honour this trust. From each group's perspective, trustees will be trustworthy towards trusters of the own group (that are encountered with a probability of α) and untrustworthy towards the outgroup (encountered with a probability of $1 - \alpha$). The probability α is common knowledge, but the group identity of the trustee is not directly observable to the truster in any particular encounter. The payoffs of both sides are ordered so to capture: (1) the benefits of a trusting truster and a trustworthy trustee engaging in interactions (e.g., long term cooperation); (2) the incentive of the untrustworthy trustee to abuse the trust given by a trusting truster; (3) the reluctance of the truster to trust given the possibility that they will encounter an untrustworthy trustee.

In Figure S1, the payoff R_I captures the high benefits of cooperation with trustworthy ingroup trustees; R_O captures the low benefits of cooperation with a trustworthy outgroup trustees; P captures the rewards of staying out of the interaction; T captures the temptation of not honouring trust; and S captures the low payoff obtained by a truster whose trust has not been honoured by the trustee. The ordering of the payoffs is $R_I > T > R_O > P > S$. We assume that $R_I > T$ and $R_O < T$, so that trustees always honour the trust of an ingroup truster and never honour the outgroup truster's trust. Thus, in this game, the truster would always want to trust an ingroup trustee and never trust the outgroup trustee. We assume that this game is played once with each player, or that players do not recognize their partners in future transactions. We also assume that trusters reveal their group identity upon giving trust to trustees.

[1] Note that in our setup, α corresponds to the share of one's ingroup in the total population. Strictly speaking, if an individual is a member of the group i that makes up share α_i of the total population, their probability of encountering their ingroup member is $(\alpha_i n - 1)/(n - 1)$, where n is the number of individuals in the total population. For simplicity, we assume that n is sufficiently large so that this probability can be approximated by α_i.

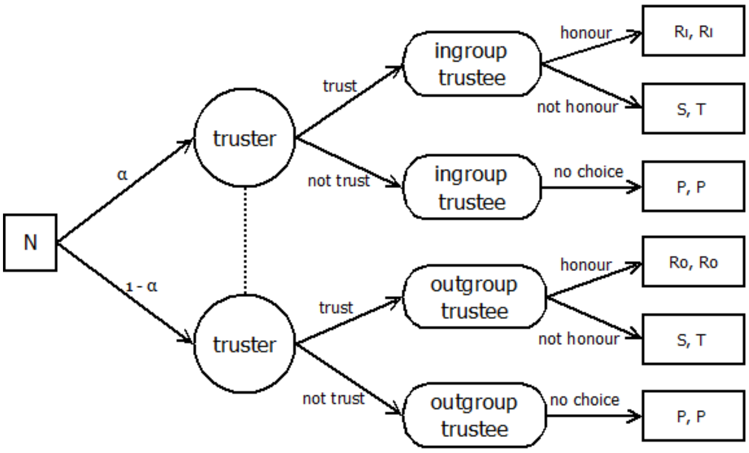


Figure S1. Trust game with incomplete information.

Given the probability of meeting a trustworthy ingroup trustee α, the truster would be willing to trust a trustee if their expected payoff of doing so is higher than the payoff of remaining outside of the interaction and not trusting at all:[2]

$$\alpha R_I + (1 - \alpha)S > P \tag{1}$$

Rearranging shows that the truster will abstain from placing trust if α is less than the threshold value:

$$\alpha^* = \frac{P - S}{R_I - S} \tag{2}$$

Following this, under the conditions where α is below this threshold ($\alpha \leq \alpha$*), a truster's expected payoff is higher when trust is not given in the first stage of the trust game (a truster does not give trust to any trustee they encounter, i.e., they unconditionally distrust). On the contrary, if α is above this threshold ($\alpha > \alpha$*), trusters trust any trustee they encounter in the first stage of the trust game (i.e., they unconditionally trust). If a trustee could reliably signal their type (here: group identity), trusters could recognize the signal and use it to identify trustworthy trustees and avoid the untrustworthy ones (trusting ingroup trustees who signal, and distrusting anyone else, i.e., conditionally trust). The introduction of signalling of one's type in the trust game can, thus, allow for the emergence of a more collectively beneficial equilibrium in which trustees signal their group identity and are given trust by ingroup trusters conditional on the trusters having observed this signal. In setups with signalling, signals are type-separating if only trustworthy trustees can afford to send them [1,2]. In our setup, we make a simplifying assumption that group members can signal their group identity accurately, but do not have an incentive, the necessary

[2] These results correspond to the results in Przepiorka and Diekmann [1], but we substitute the terms and notation of long- and short-term trustees to in- and out-group trustees to align it with our particular context.

knowledge, or the possibility to imitate signals of the outgroup. This could be either because it is cheaper to send the signal of one's own group than to imitate the one of the outgroup (e.g., a difficult to fake accent), or because either the benefits of cooperating with the outgroup or the benefits of abusing the trust given by the outgroup are not sufficiently high to offset the costs of signalling the outgroup's identity.[3] In Section S2.1, we discuss the consequences of relaxing this assumption.

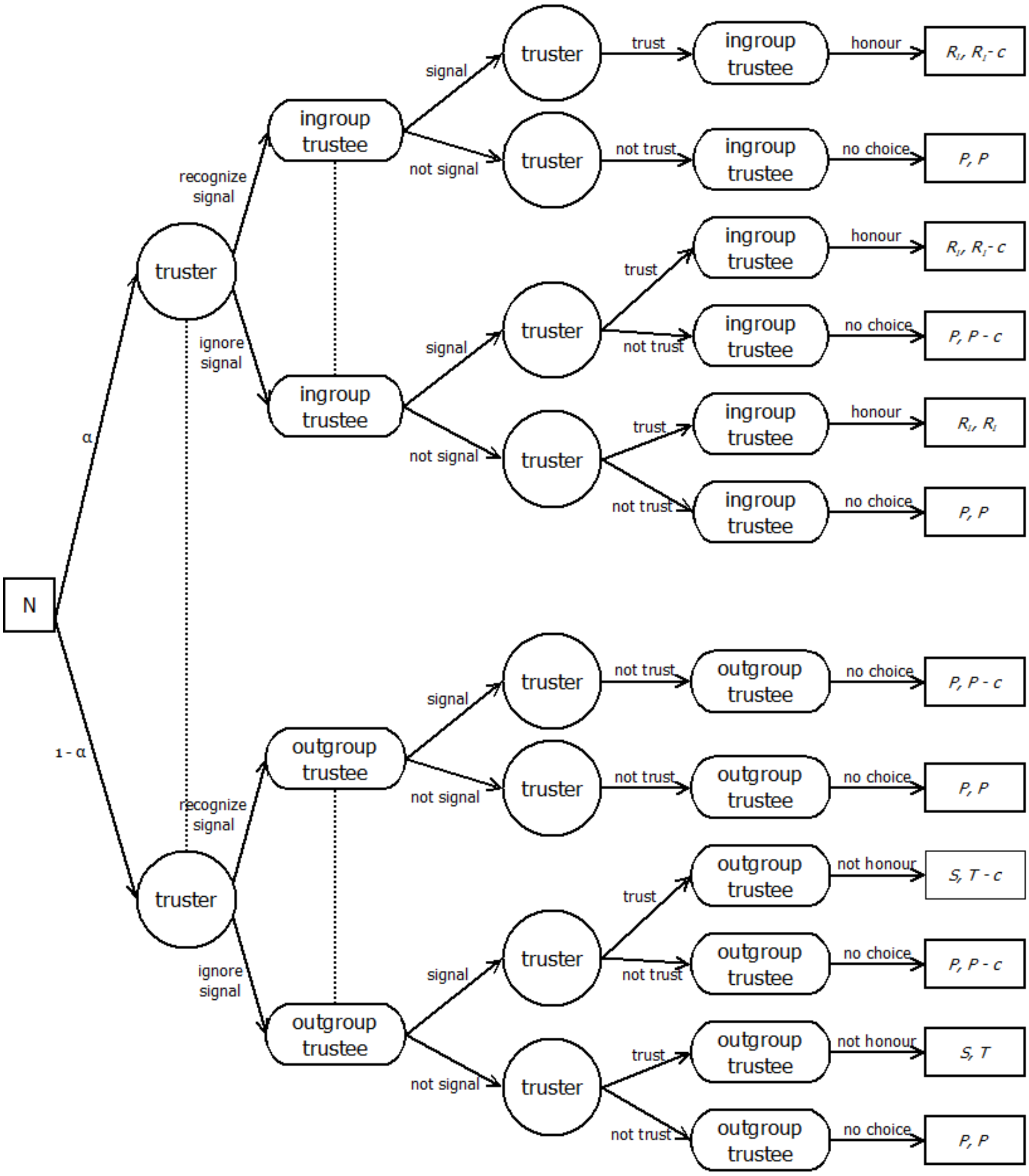


Figure S2. Simplified trust game with signalling by trustees and signal recognition by trusters. In this game, trustees can only send accurate signals of their group identity (they cannot fake signals of the outgroup) at cost *c* and,

[3] For a detailed discussion of the conditions under which this is the case, see Przepiorka and Diekmann [1].

upon giving trust, the truster reveals their group identity. Once the truster reveals their identity, the trustee will always honour the ingroup truster's trust ($R_I > T$) and abuse the outgroup truster's trust ($R_O < T$).

Under these assumptions, if a truster trusts an ingroup trustee, they both obtain payoff R_I and if they trust an outgroup trustee, the truster always obtains S and the trustee T. In Figure S2 we show the trust game with incomplete information and signalling including the simplifying assumptions listed above.

Assumption 1. Two groups co-exist in a population, with each constituting a share α_i of the population, where $i \in \{1, 2\}$ and $\alpha_1 \leq \alpha_2$.

For different α_2 and α^*, there are three scenarios possible:

1. Scenario where $\alpha_2 \leq \alpha^*$. In the absence of signalling, trusters are unconditionally distrusting;
2. Scenario where $\alpha_1 \leq \alpha^*$ and $\alpha_2 > \alpha^*$. In the absence of signalling, trusters are unconditionally distrusting in the former, and unconditionally trusting in the latter group.
3. Scenario where $\alpha_1 > \alpha^*$. In the absence of signalling, trusters are unconditionally trusting.

To find the conditions under which trustees signal their identity and trusters condition their trust upon observing these signals (henceforth: conditional trusting), we can rely on the insights about the truster strategies under different α parameters as per the three scenarios above (also, see Eq. 2). In the absence of signalling by trustees, if $\alpha \leq \alpha^*$, trusters unconditionally distrust; whereas if $\alpha > \alpha^*$, trusters unconditionally trust. We assume that trusters do not recognize signals by default; recognizing signals needs to pay off against other available strategies. We make this assumption because signals sent by one agent are only valuable to the extent that they mean something to another agent. For example, a certain tattoo can be a reliable marker of one's group identity only if the group members are aware of its meaning. Consequently, the expected payoffs trusters obtain from conditionally trusting – compared to unconditionally (dis)trusting will be dependent on the share of signalling trustees in their own group. Similarly, the decision on the trustee side as to whether to display (i.e., send) signals depends on the existence of conditionally trusting trusters in their group.

Whereas agents' payoffs depend on the strategies of the outgroup (e.g., whether a trustee is given trust by the outgroup), their best responses only depend on the ingroup agents' strategies and are independent of the exact strategy of the outgroup agents. That is, an agent does not have to condition their signalling decision nor their trusting decision on the exact strategies of the outgroup. For instance, regardless of whether the outgroup trusters conditionally trust, a trustee's decision to signal or not is only determined in relation to the conditional trusting of the ingroup trustees. This simplifies our analysis considerably (but see Section S2).

S1.1. Signalling and non-signalling equilibria

We first consider the signalling norm equilibrium.

Theorem 1 (Signalling strategy theorem). There exists a Nash Equilibrium where trusters condition their trusting on signals of the ingroup and trustees display group identity signals (signalling equilibrium) if and only if the signalling costs are offset by the benefits of parochial cooperation established with the help of signalling ($c \leq \alpha[R_I - P]$, where $R_I - P$ captures the benefits of cooperation).

Proof 1. In a group where all trusters are conditionally trusting, signalling is the best response for a trustee as long as the benefits of parochial cooperation offset the costs of sending the signal:

$$\alpha(R_I - c) + (1 - \alpha)(X - c) \geq \alpha P + (1 - \alpha)X \tag{3}$$

In Eq. 3, X denotes the payoff obtained by the trustee in an encounter with the outgroup truster. That is, if the trustee faces: (1) an outgroup with $\alpha > \alpha^*$ in a non-signalling equilibrium, the outgroup truster will be unconditionally trusting and the payoff $X = T$; (2) an outgroup with $\alpha \leq \alpha^*$ in a non-signalling equilibrium, the outgroup truster will be unconditionally distrusting and the payoff $X = P$; (3) an outgroup with any α that is in a signalling equilibrium, the outgroup trustee will be conditionally trusting and disregard any behaviour that does not send their ingroup signal, so the payoff $X = P$. In Eq. 3, the elements $(1 - \alpha)X$ on both sides of the inequality cancel out, which is why the strategy of outgroup trusters does not enter the best response consideration of the ingroup trustee. We will continue all of our analyses by using X as the placeholder for payoffs dependent on the strategies of outgroup members. Simplifying Eq. 3, we obtain the condition for signal sending by trustees:

$$c \leq \alpha(R_I - P) \tag{4}$$

When all trustees are signalling, conditional trusting is the best response for a trustee in a group with $\alpha \leq \alpha^*$ if the expected payoff of parochial cooperation with the signalling ingroup is higher than the payoff from unconditional distrusting:

$$\alpha R_I + (1 - \alpha)P \geq \alpha P + (1 - \alpha)P \tag{5}$$

The above inequality always holds, as $R_I > P$ in our setup. Conditional trusting is the best response for a trustee in a group with $\alpha > \alpha^*$ if the expected payoff of parochial cooperation with the signalling ingroup is higher than the payoff from unconditional trusting:

$$\alpha R_I + (1 - \alpha)P \geq \alpha R_I + (1 - \alpha)S \tag{6}$$

The above inequality also always holds, as $P > S$ in our setup.

Theorem 2. In a group with $\alpha \leq \alpha^*$, there exists a Nash Equilibrium where trusters unconditionally distrust all trustees and trustees do not display group identity signals (non-signalling equilibrium). In a group with $\alpha > \alpha^*$, there exists a non-signalling equilibrium where trusters unconditionally trust all trustees and trustees do not display group identity signals.

Proof 2. In a group *with* $\alpha \leq \alpha^*$ where all trusters are unconditionally distrusting, not signalling is the best response for a trustee when:

$$\alpha P + (1 - \alpha)X \geq \alpha(P - c) + (1 - \alpha)(X - c) \tag{7}$$

Which is always the case unless $c < 0$. We do not consider any cases where $c < 0$, that is, the act of signalling nets benefits for trustees in itself. In a group *with* $\alpha > \alpha^*$ where all trusters are unconditionally trusting, the condition for trustees non-signalling strategy is the same as in Eq. 7, but the benefits from encountering the ingroup are R_I rather than P.

In a group with $\alpha \leq \alpha^*$ where no trustees signal their ingroup identity, unconditional distrust is always the best response for trusters, as there are no benefits from switching to conditional trust in their group (they will always obtain the payoff P, regardless of whether they meet the ingroup or the outgroup). In a group with $\alpha > \alpha^*$ where no trustees signal their ingroup identity, unconditional trust is the best response for trusters as long as its benefits are higher than those of conditional trusting:

$$\alpha R_I + (1 - \alpha)S \geq P \tag{8}$$

The inequality above is equivalent to Eq. 1 and Eq. 2. Hence, it holds by definition if $\alpha > \alpha^*$.

S1.2. Shifting to the signalling equilibrium

We describe the emergence of the signalling norm within a group as a shift from a non-signalling equilibrium in which no trustees signal their group identity and no trusters condition their trust on having recognized the signal to a signalling equilibrium in which all trustees signal and all trusters conditionally trust. To understand these dynamics, we evaluate a situation in which a few individuals deviate from best response strategies in the non-signalling state. Note that we only consider a signalling norm to have emerged if trustees signal their identity *and* trusters recognize these signals as per Theorem 1.

In our analyses, we denote the share of conditionally trusting trusters in one's group as γ (therefore, the share of the best response strategy in the non-signalling equilibrium is $(1 - \gamma)$). We denote the share of conditionally trusting trusters in the outgroup as δ (therefore, the share of the best response outgroup strategy in the non-signalling equilibrium $(1 - \delta)$). The exact payoffs will differ based on the scenarios as discussed earlier in this section; for brevity, we only focus on Scenario 2, where $\alpha_1 \leq \alpha^*$ and $\alpha_2 > \alpha^*$.[4] We start by considering a decision of a trustee to signal their group identity in a non-signalling state, but once some trusters have deviated from the best response strategy in their equilibrium. Recall from our previous analyses that the trustees in the group $\alpha \leq \alpha^*$ will be facing unconditionally distrusting ingroup trusters and unconditionally trusting outgroup trusters (while the opposite holds for the group with $\alpha > \alpha^*$).

We begin by looking at the decision of the members of the group with $\alpha \leq \alpha^*$. A trustee in this group signals their identity if the benefits from encountering conditionally trusting ingroup trusters and the gains from abusing the trust of the unconditionally trusting outgroup trusters offset the costs of signalling:

$$\alpha[\gamma(R_I - c) + (1-\gamma)(P - c)] + (1-\alpha)[\,\delta(P - c) + (1-\delta)(T - c)] \geq \alpha[\gamma P + (1-\gamma)P] + (1-\alpha)[\,\delta P + (1-\delta)T] \tag{9}$$

Rearranged, this inequality implies that a trustee benefits from signalling if the share of ingroup trusters who have started to conditionally trust is as follows:

$$\gamma^* \geq \frac{c}{\alpha(R_I - P)} \tag{10}$$

Now, we move on to determine the conditions under which trusters in the same group will prefer conditional trusting to unconditional distrusting. We define the share of signalling trustees in a group with β (therefore, the share of non-signalling trustees in the same group will be $(1 - \beta)$. In our setup, conditionally trusting trusters only give trust to signalling ingroup trustees. That is, trusters do not consider outgroup signals as relevant. Therefore, the share of signalling outgroup trustees is irrelevant for a truster's decision whether to conditionally trust or not, and we do not consider it here (but see Section S2.2.). A truster conditionally trusts if the expected payoff from cooperating with signalling ingroup trustees and not trusting anyone else outweighs the expected payoff from unconditional distrust:

$$\alpha[\beta R_I + (1-\beta)P] + (1-\alpha)P \geq P \tag{11}$$

That is, trusters are indifferent between conditional trust and unconditional distrust when there are no signalling ingroup members – and prefer conditional trust when there is a positive number of signalling ingroup members. Therefore, trusters conditionally trust when:

$$\beta^* \geq 0 \tag{12}$$

Now, we look at the decisions of individuals in the group with $\alpha > \alpha^*$. The analysis is similar to the one in Eq. 9. We once again denote the share of conditionally trusting trusters in one's group as γ

[4] This is also the scenario we focus on in our agent-based simulations.

(yet, as there are no unconditionally distrusting trusters in this group, the share of unconditionally trusting trusters is (1 – γ)). We denote the share of conditionally trusting trusters in the outgroup as δ (therefore, the share of the unconditionally distrusting outgroup trusters is (1 – δ)). A trustee signals their identity if there is a sufficient number of conditional trusters in their group to make signalling worthwhile:

$$\begin{aligned}\alpha[\gamma(R_I - c) + (1-\gamma)(R_I - c)] + (1-\alpha)[\,\delta(P-c) + (1-\delta)(P-c)] \\ \geq\ \alpha[\gamma P + (1-\gamma)R_I] + (1-\alpha)[\,\delta P + (1-\delta)P]\end{aligned} \tag{13}$$

This inequality simplifies to Eq. 10.

Finally, we analyse conditions under which trusters in the group with $\alpha_2 > \alpha^*$ conditionally trust. Once again, we denote the share of signalling trustees in a group with β (therefore, the share of non-signalling trustees in the same group will be (1 – β); the share of signalling trustees in the outgroup is irrelevant). Trusters conditionally trust if sufficiently many ingroup agents signal and the benefits of avoiding the exploitation by the outgroup are not outweighed by the losses stemming from not trusting the non-signalling ingroup:

$$\alpha[\beta R_I + (1-\beta)P] + (1-\alpha)P \geq\ \alpha R_I + (1-\alpha)S \tag{14}$$

This inequality holds when the share of signalling ingroup trustees is as below:

$$\beta^* \geq 1 - \frac{(1-\alpha)(P-S)}{\alpha(R_I - P)} \tag{15}$$

This threshold shows the payoff ratio of not trusting versus trusting an outgroup member [(1 – α)(P – S)] to trusting versus not trusting an ingroup member [α(R_I – P)].

Regardless of the exact scenario, the condition in Eq. 10 for both groups, the condition in Eq. 12 for groups with $\alpha \leq \alpha^*$ and the condition in Eq. 15 for groups with $\alpha > \alpha^*$ remain the same. In Table S1 below, we summarize the conditions for switching to signalling and conditionally trusting strategies if some individuals in the population start using strategies that deviate from the non-signalling equilibrium best responses. These conditions show that the threshold β_i^* for switching to conditional trusting is lower for trusters in a group with $\alpha \leq \alpha^*$ than for those in a group with $\alpha > \alpha^*$. That is, trusters in the former group who start conditionally trusting can remain doing so without incurring any losses; while those in the latter group need a sufficiently high number of signalling trustees before adopting the conditionally trusting strategy. This, in turn, ensures that the threshold γ_i^* for trustees to switch to signalling is more easily passed in groups with $\alpha \leq \alpha^*$ than in those with $\alpha > \alpha^*$. Given this, we expect that groups with $\alpha \leq \alpha^*$ will always be able to shift into the signalling equilibrium as long as the random strategy changes occur with sufficient frequency. On the other hand, we expect groups with $\alpha > \alpha^*$ to be less likely to successfully shift to the signalling equilibrium.

Situation	γ_i*	β_i*
$\alpha_i \leq \alpha^*$	$\gamma \geq \frac{c}{\alpha(R_I - P)}$	$\beta \geq 0$
$\alpha_i > \alpha^*$		$\beta \geq 1 - \frac{(1-\alpha)(P-S)}{\alpha(R_I - P)}$

Table S1. Conditions for switching from the non-signalling equilibrium best response strategies. Share γ^* refers to the share of conditionally trusting trusters in one's ingroup that are necessary before a trustee switches from a non-signalling to a signalling strategy. Share β^* refers to the share of signalling trustees in one's group that are necessary for a trusters to switch to the conditionally trusting strategy (from either unconditional trust or unconditional distrust).

S2. Model extensions

In this section, we discuss how (individually) relaxing each one of the three main assumptions in our model impacts our results.

S2.1. Dishonest signalling

We first analyse the consequences of relaxing the assumption that trustees can only send ingroup signals (i.e., signals corresponding to their group identity). As in our original setup, trustees have to decide on their signalling strategy before encountering a particular truster. That is, trustees cannot condition their signalling behaviour on the truster's group identity, as it is only revealed once trust is given by the truster. In this setup, trustees can: (1) not signal their group identity; (2) send a signal of their true group identity (i.e., ingroup signal); (3) send a signal of the outgroup identity (i.e., outgroup signal, whereby they "pretend" to be an outgroup trustee when signalling, but act according to their actual group identity if given trust). We analyse the most adverse setting, where sending an outgroup signal is as costly as sending an ingroup one. In this setup, one could benefit from sending an outgroup signal if outgroup trustees are conditionally trusting. While signalling outgroup identity blocks one from being given trust by the ingroup, it allows them to exploit the conditionally trusting outgroup. Given this, we now have to discuss equilibria bearing both groups' strategies in mind. First, we start by showing that there is no dishonest signalling equilibrium where both groups send dishonest signals and conditionally trust.

Theorem 3. There is no Nash Equilibrium where trusters in both groups condition their trusting on signals of the ingroup and trustees in both groups display outgroup identity signals.

Proof 3. If all trustees in the ingroup and the outgroup send outgroup signals, trusters in a group with $\alpha \leq \alpha^*$ will only conditionally trust if that brings more benefit than unconditionally distrusting, which is the case when:

$$\alpha P + (1 - \alpha)S \geq \alpha P + (1 - \alpha)P \tag{16}$$

This condition is never satisfied, as $P > S$ in our setup. Similarly, trusters in a group with $\alpha > \alpha^*$ will only conditionally trust if that brings more benefit than unconditionally trusting, which is the case when:

$$\alpha P + (1 - \alpha)S \geq \alpha R_I + (1 - \alpha)S \tag{17}$$

This condition is also never satisfied, as $R_I > P$ in our setup. As trusters in either group have no incentive to conditionally trust, trustees will never send outgroup signals as long as $c > 0$.

Next, we show that there is a Nash Equilibrium where trustees in both groups send ingroup signals and trusters recognize these signals, despite the fact that there is an opportunity to send signals imitating the outgroup.

Theorem 4. There is a Nash Equilibrium where trusters in both groups condition their trusting on signals of the ingroup and trustees in both groups display ingroup group identity signals if the signalling costs are offset by the benefits of parochial cooperation established with the help of signalling ($c \leq \alpha[R_I - P]$) and the ingroup is sufficiently large as to offset the temptation to send outgroup signals to the conditionally trusting outgroup ($\alpha \geq [T - P]/[R_I - 2P + T]$).

Proof 4. Similar to our analyses in Proof 1, we can show that trusters in both groups will always conditionally trust as long as all trustees are sending group identity signals. We then need to compare the utility of sending an ingroup versus sending an outgroup signal for trustees in both groups.

Assuming both signals are equally costly to send, ingroup signals will be preferred once: (1) the benefits of parochial cooperation minus the costs of being distrusted by the outgroup are higher than (2) the benefits of abusing the conditionally trusting outgroup minus the losses of being distrusted by the ingroup (when sending outgroup signals). This comparison yields the following condition for sending ingroup signals:

$$\alpha \geq \frac{T - P}{R_I - 2P + T} \tag{18}$$

Compared to our baseline analyses, the condition on the value of α under which ingroup signalling is beneficial for a group is more constrained. This suggests, that, if there is a possibility to send the outgroup signal, groups with an α below the threshold in Eq. 18 would prefer to send outgroup, rather than ingroup signals. On the other hand, as the benefit of parochial cooperation R_I increases, the groups that prefer sending ingroup signals can become smaller. We further show that there is one equilibrium in which one group's trustees can send outgroup signals and exploit the outgroup trusters if the condition in Eq. 18 is not satisfied for one of the groups.

Theorem 5. There is a Nash Equilibrium where trusters in a group with $\alpha \leq \alpha^*$ condition their trusting on signals of the ingroup and trustees display outgroup identity signals if the cost of sending signals is sufficiently small to be offset by the exploitation of the conditionally trusting outgroup ($c \leq \alpha[P - T] - P + T$) and the ingroup is sufficiently small so that outgroup signals are preferred to ingroup ones ($\alpha < [T - P]/[R_I - 2P + T]$). In this equilibrium, trusters in the outgroup with $\alpha \geq \alpha^*$ condition their trusting on signals of the ingroup and trustees display honest group identity signals if the ingroup is sufficiently large to offset the losses from trusters being exploited by the outgroup and ingroup signals are preferred to outgroup ones ($\alpha \geq [T - P]/[R_I - 2P + T]$).

Proof 5. We can show that, if the condition in Eq. 18 is not satisfied for the group with $\alpha \leq \alpha^*$, trusters in this group will always conditionally trust and trustees always send outgroup signals. When facing a group where trustees send outgroup signals, trusters in the group with $\alpha > \alpha^*$ will conditionally trust and trustees will always send ingroup signals as long as the cost condition in Eq. 4 and the ingroup signalling condition in Eq. 18 are both satisfied.

S2.2. Distrusting outgroup signals

We now relax the assumption that trusters cannot purposefully discriminate against the signalling outgroup trustees. To do so, we introduce an additional strategy of conditional distrust, where trusters trust all trustees unless they have observed an outgroup signal. That is, now a truster can pursue one of the four different strategies: (1) trust all trustees (unconditional trust); (2) distrust all trustees (unconditional distrust); (3) trust only trustees who display ingroup signal (conditional trust); (4) distrust only trustees who display outgroup signal (conditional distrust).

We proceed to show that, even in the presence of conditional distrust, signalling and signal recognition can be equilibrium strategies under certain conditions. Here, we focus on key insights only, rather than exploring all of the possible equilibria. As the outgroup is able to condition their trusting behaviour on the ingroup's signalling, we need to discuss equilibria bearing both groups' strategies in mind.

Theorem 6. There exists a Nash Equilibrium where trusters condition their trusting on signals of the ingroup and trustees display group identity signals if the signalling costs are offset by the benefits of parochial cooperation established with the help of signalling ($c \leq \alpha[R_I - P]$) and the outgroup is also in a signalling equilibrium.

Proof 6. If both groups are in a signalling equilibrium, trustees will be facing conditionally trusting ingroup and outgroup trusters. In this case, signalling is the best response for trustees following the same cost condition as in Eq. 4. As long as the ingroup trustees are signalling, conditional trusting is preferred by trusters to unconditional trusting, unconditional distrusting, and conditional distrusting. The same holds for both groups, and the rest of the analysis parallels our main analysis in Proof 1. However, it is more interesting to understand whether signalling and signal recognition can be equilibrium strategies for groups facing an outgroup that is not in a signalling equilibrium.

Theorem 7. There exists a Nash Equilibrium for a group with $\alpha \leq \alpha^*$ facing a group with $\alpha > \alpha^*$ where, in the former group, trusters condition their trusting on signals of the ingroup and trustees display group identity signals if the signalling costs are offset by the benefits of parochial cooperation and *offset the risks of being discriminated against by the outgroup ($c \leq \alpha[R_I - 2P + T] + P - T$). In this equilibrium, trusters in the group with $\alpha > \alpha^*$ trusters condition their distrusting on signals of the outgroup and trustees do not display group identity signals.*

Proof 7. In a group with $\alpha \leq \alpha^*$, if trusters are conditionally trusting, and the outgroup trusters are conditionally distrusting, trustees will signal their identity if the benefits of parochial cooperation, minus the costs of being discriminated against the outgroup, offset the benefits of not signalling and therefore being distrusted by the conditionally trusting ingroup but trusted by the conditionally distrusting outgroup:

$$\alpha(R_I - c) + (1 - \alpha)(P - c) \geq \alpha P + (1 - \alpha)T \tag{19}$$

Simplifying this condition yields the following signalling cost condition:

$$c \leq \alpha(R_I - 2P + T) + P - T \tag{20}$$

Compared to the cost condition in the main analysis (Eq. 4), the cost that trustees are willing to bear in this setup is lower. That is, introducing a possibility of conditional distrusting by the outgroup lowers the willingness to pay for costly signals.

If all ingroup trustees signal and no outgroup trustees signal, trusters will always prefer conditional trusting in the group with $\alpha \leq \alpha^*$. In the group with $\alpha > \alpha^*$, if the ingroup trustees are not signalling, but the outgroup trustees are, conditional distrust is the best response if it brings as much or more benefit compared to unconditional trusting, which is the case when:

$$\alpha R_I + (1 - \alpha)P \geq \alpha R_I + (1 - \alpha)S \tag{21}$$

This is always the case as $P > S$ in our setup. In this group, trustees will always prefer not to signal, given that trusters conditionally trust, unless $c > 0$.

Theorem 8. There exists a Nash Equilibrium for a group with $\alpha > \alpha^*$ facing a group with $\alpha \leq \alpha^*$ where trusters condition their trusting on signals of the ingroup and trustees display group identity signals if the signalling costs are offset by the benefits of parochial cooperation ($c \leq \alpha[R_I - P]$). In this equilibrium, trusters in the group with $\alpha \leq \alpha^*$ condition their distrusting on signals of the outgroup and trustees do not display group identity signals.

Proof 8. In a group with $\alpha > \alpha^*$, if trusters are conditionally trusting, and the outgroup trusters are unconditionally distrusting, trustees will signal their identity if the benefits of parochial cooperation offset the signalling costs, with the same condition as in Eq. 4. If ingroup trustees are signalling and the outgroup trustees are not, trusters will conditionally trust if this brings more benefits than unconditional trust, which is the condition in Eq. 6 (main analysis) and always holds in our setup.

For the group with $\alpha \leq \alpha^*$ facing conditionally distrusting ingroup trusters (who will always trust the ingroup) and conditionally trusting outgroup trusters (who will never trust the outgroup) not signalling is preferred to signalling if:

$$\alpha R_I + (1-\alpha)P \geq \alpha(R_I - c) + (1-\alpha)(P - c) \tag{22}$$

This is always the case unless $c < 0$. In this group, trusters will conditionally distrust if this brings them as much or more benefits than unconditionally distrusting, which is the case when:

$$\alpha R_I + (1-\alpha)P \geq P \tag{23}$$

This condition always holds as $R_I > P$ in our setup.

Given our conclusions from the main analysis that costly signals are more likely to evolve in groups with $\alpha \leq \alpha^*$ than in groups with $\alpha > \alpha^*$, we can expect that a population starting in a non-signalling equilibrium will be more likely to end up in the equilibrium described in Theorem 7 than in those described in Theorems 6 and 8.

S2.3. Noise in trustee perception of truster group identity

In our model, trustees decide to signal their identity before being aware of the truster's identity. Here, we relax the assumption that trusters' identity is always accurately perceived by trustees once trust is given. We analyse the amount of noise that can be present in trustees' perception of the truster type before signalling and signal recognition stop being equilibrium strategies. With a probability ζ, the trustee will misinterpret the truster's identity, thereby dishonouring the trust given by the ingroup (and obtain T, with the truster obtaining S), and honouring trust given by the outgroup trustee (whereby both obtain R_O). With a probability $(1 - \zeta)$ the trustee will correctly interpret the truster's group identity and correspondingly honour or dishonour the given trust. However, we constrain ζ so that $0 \leq \zeta \leq 0.5$. Within this range of values, the correlation between the group identity information revealed by the truster (or perceived by the trustee) when placing trust and the truster's actual group identity is positive; when $\zeta = 0.5$, there is no correlation between the two whatsoever. We do not consider cases when $\zeta > 0.5$, since this would imply that there is a negative correlation between the cues that the truster reveals about themselves upon placing trust and their actual group identity. In our analysis, we do not consider the possibility that trustees might have other assumptions regarding the likelihood of trusters from different groups giving or not giving them trust.

We first re-evaluate the condition in Eq. 1 to determine under which conditions trusters would be willing to unconditionally trust rather than unconditionally distrust trustees, knowing that trustees might misinterpret trusters' group identities with a probability ζ:

$$\alpha[(1-\zeta)R_I + \zeta S] + (1-\alpha)[(1-\zeta)S + \zeta R_O] \geq P \tag{24}$$

Rearranging Eq. 24 for α, we can see that it simplifies to Eq. 2 if there is no noise ($\zeta = 0$):

$$\alpha \geq \frac{P - S - \zeta(R_O - S)}{R_I - S - \zeta(R_I + R_O - 2S)} \tag{25}$$

Rearranging Eq. 24 for noise parameter ζ, we get the following condition:

$$\zeta^* \leq \frac{\alpha(R_I - S) + S - P}{\alpha(R_I - 2S + R_O) + S - R_O} \tag{26}$$

As long as α and ζ are such that Eq. 25 (and 26) are satisfied, trusters with $\alpha \leq \alpha^*$ will unconditionally distrust, whereas trusters with $\alpha > \alpha^*$ will unconditionally trust all trustees in the absence of signalling. We will proceed to only analyse this situation, and defer further evaluation of equilibria when $\zeta > \zeta^*$ to future research.

Theorem 9. There exists a Nash Equilibrium where trusters condition their trusting on signals of the ingroup and trustees display group identity signals (signalling equilibrium) if and only if the signalling costs are offset by the benefits of parochial cooperation established with the help of signalling ($c \leq \alpha[R_I - P] - \zeta[R_I - T]$) and noise is sufficiently low ($\zeta \leq \zeta^*$ and $\zeta \leq [R_I - P]/[R_I - S]$ when $\alpha \leq \alpha^*$ or $\zeta \leq [P - S]/[R_O - S]$ when $\alpha > \alpha^*$).

Proof 9. In a group where all trusters are conditionally trusting, and trustees erroneously perceive trusters group identity with probability ζ, signalling is the best response for a trustee as long as the benefits of parochial cooperation offset the costs of sending the signal:

$$\alpha[(1-\zeta)(R_I - \mathrm{c}) + \zeta(T - c)] + (1-\alpha)(X - c) \geq \alpha P + (1-\alpha)X \tag{27}$$

In Eq. 27, X denotes the payoff obtained by the trustee in an encounter with the outgroup truster. That is, if the trustee faces: (1) an outgroup with $\alpha > \alpha^*$ in a non-signalling equilibrium, the outgroup truster will be unconditionally trusting and the payoff $X = (1 - \zeta)T + \zeta R_O$; (2) an outgroup with $\alpha \leq \alpha^*$ in a non-signalling equilibrium, the outgroup truster will be unconditionally distrusting and the payoff $X = P$; (3) an outgroup with any α that is in a signalling equilibrium, the outgroup trustee will be conditionally trusting and distrust any trustee that does not send their ingroup signal, so the payoff $X = P$. In Eq. 3, the elements $(1-\alpha)X$ on both sides of the inequality cancel out. Simplifying Eq. 27, we obtain the condition for signal sending by trustees:

$$c \leq \alpha(R_I - P) - \alpha\zeta(R_I - T) \tag{28}$$

Eq. 28 sets the condition for signalling costs trustees are willing to bear. Here, $R_I - P$ captures the benefits of cooperation and $R_I - T$ the cost of noise in trustees perception of truster identity. The higher the noise ζ, holding everything constant, the lower the cost trustees are willing to bear. Finally, when there is no noise (ζ = 0), Eq. 28 simplifies to the main cost condition in Eq. 4.

When all trustees are signalling, conditional trusting is the best response for a truster in a group with $\alpha \leq \alpha^*$ if the expected payoff of parochial cooperation with the signalling ingroup is higher than the payoff from unconditional distrusting, accounting for noise in trustee's perception of truster group identity:

$$\alpha[(1-\zeta)R_I + \zeta S] + (1-\alpha)P \geq \ \alpha P + (1-\ \alpha)P \tag{29}$$

Simplifying this inequality yields the condition for maximum noise ζ that trustees with $\alpha \leq \alpha^*$ are willing to tolerate upon conditionally trusting:

$$\zeta \ \leq \frac{R_I - P}{R_I - S} \tag{30}$$

Similarly, conditional trusting is the best response for a truster in a group with $\alpha > \alpha^*$ if the expected payoff of parochial cooperation with the signalling ingroup is higher than the payoff from unconditional trusting, accounting for noise in trustee's perception of truster group identity:

$$\alpha[(1-\zeta)R_I + \zeta S] + (1-\alpha)P \geq \ \alpha[(1-\zeta)R_I + \zeta S] + (1-\ \alpha)[(1-\zeta)S + \zeta R_O] \tag{31}$$

Simplifying Eq. 31 yields the condition for maximum noise ζ that trustees with $\alpha > \alpha^*$ are willing to tolerate upon conditionally trusting:

$$\zeta \ \leq \frac{P - S}{R_O - S} \tag{32}$$

Theorem 2. In a group with $\alpha \leq \alpha^*$ there exists a Nash Equilibrium where trusters unconditionally distrust all trustees and trustees do not display group identity signals (non-signalling equilibrium). In a group with $\alpha > \alpha^*$ there exists a non-signalling equilibrium where trusters unconditionally trust all trustees and trustees do not display group identity signals. The existence of this equilibrium is conditional on the level of noise being sufficiently low ($\zeta \leq \zeta^*$).

Proof 2. We still assume that Eq. 26 is satisfied. If $\alpha \leq \alpha^*$ all trusters are unconditionally distrusting, not signalling is the best response for a trustee when:

$$\alpha P + (1-\alpha)[(1-\zeta)T + \zeta R_O] \geq \ \alpha(P-c) + (1-\ \alpha)[(1-\zeta)(T-c) + \zeta(R_O - c)] \tag{33}$$

Which is always the case unless $c < 0$. For trusters in this group, unconditional distrust is always the best response, as there are no benefits from switching to conditional trust in their group (they will always obtain the payoff P, regardless of whether they meet the ingroup or the outgroup).

Similarly, in a group *with* $\alpha > \alpha^*$ where all trusters are unconditionally trusting, trustees never benefit from signalling unless $c < 0$. If $\alpha > \alpha^*$ where no trustees signal their ingroup identity, unconditional trust is the best response for trusters as long as its benefits are higher than those of conditional trusting:

$$\alpha[(1-\zeta)R_I + \zeta S] + (1-\alpha)[(1-\zeta)S + \zeta R_O] \geq \ P \tag{34}$$

Note that this inequality is equivalent to Eq. 24, and is automatically satisfied if Eq. 26 is satisfied.

S3. Agent-based model details

S3.1. Social learning

We build our agent-based model in NetLogo [3]. At the beginning of each round, each agent is randomly assigned into a role of a truster or a trustee. They are matched with another agent with a different role. The probability of encountering an ingroup member depends on the share of the ingroup in the population (p_i). Additionally, we inspect the interaction between the group's share in the population and the likelihood that individuals interact within their own group (ingroup assortment φ) to study the interaction between relative group size and signalling norm emergence dynamics. The interaction is played out based on each agent's group identity and traits, with the outcomes depending on agents' group identities and their strategies (outcomes of different strategies encountering each other are as per Figure S2). At the end of each round, an agent's payoff from that round is added to the list of their payoffs from up to twenty previous rounds; this list is then weighted to discount older payoffs and summed into their fitness score (w) over these rounds. We apply the moving average calculation as per Macy and Skvoretz [4] to obtain averaged fitness scores:

$$w_{it} = \frac{\sum_{t'=1}^{t} 0.9^{t-t'} \boldsymbol{O}_{t'}}{\sum_{t'=1}^{t} 0.9^{t-t'}} \tag{35}$$

Where w_{it} is the fitness score of agent i at time t and $\boldsymbol{O}_{t'}$ is the vector of agent's payoffs in time periods from $t' = 1$ to t. We use a weighting factor of $0.9^{t-t'}$ that assigns greater weight to more recent outcomes. For instance, the most recent payoff will be given a weight of 1, the second most recent one of 0.9, the third most recent one of 0.81, and so on.

Each agent is then given a chance to update their traits through a social learning process. Each agent picks a random other agent from the same group and decides whether to imitate both of that agent's traits. We operationalize our learning mechanism using the proportional imitation rule as per Chica and colleagues [5]: an agent *i* with a strategy *s(i)* will adopt the strategy *s(j)* of agent *j* with a probability that increases in the difference between their payoffs in the previous round and is normalized by the difference of the maximum and minimum possible payoffs between two arbitrary agents (η):

$$prob_{s(i)}^{t} s(j) = \frac{\max\{0, w_j^{t-1} - w_i^{t-1}\}}{\eta} \tag{36}$$

Combined, our weighting and learning rules present two features that have been shown to support successful learning in dynamic environments: (1) social learning through copying more successful strategies employed by others in the population; and (2) discounting older information more than recent information [6]. We allow for random mutations (probability m_o) of agents' traits at the end of every round and after each strategy updating step; if an agent is selected for mutation, either of their traits has an equal chance to be randomly changed with a probability of m_o. According to our model with payoff settings as per Figure 1 and the formula derived in Eq. 2, α^* = 0.43.

S3.2. Details on p_i and φ parameters

We run each simulation for 20 000 rounds, letting agents learn and mutate after each round. We vary the share of group 1 in the population $p_1 \in \{0.1, 0.2, 0.3, 0.4\}$. Correspondingly, this sets the share of group 2 in the population as $p_2 \in \{0.6, 0.7, 0.8, 0.9\}$. In one simulation run, we are able to evaluate two scenarios, one with group 1 facing $\alpha_1 \leq \alpha^*$ and, simultaneously, one with group 2 facing $\alpha_2 > \alpha^*$. For instance, if $p_1 = 0.4$, $p_2 = 0.6$ by the default, and the groups' respective α values will be 0.4 and 0.6. Recall from Subsection S2.1 that $\alpha^* = 0.43$ in our simulations, which means that our group 1 will always be a "minority" group with $\alpha_1 \leq \alpha^*$ and group 2 will always be a "majority" group with $\alpha_2 > \alpha^*$. Therefore, our varying of parameter p_i as above allows us to test $\alpha \in \{0.1, 0.2, 0.3, 0.4, 0.6, 0.7, 0.8, 0.9\}$.

In addition, we survey the interaction of groups' relative shares in the population with the likelihood of members interacting with their ingroup. We capture this using the parameter φ as follows: agents from groups 1 and 2 can find themselves at either location L_1 or L_2, with the former corresponding to an imaginary "habitat" of group 1 and the latter to an imaginary "habitat" of group 2. Parameter φ captures the probability that agent from group i inhabits habitat L_i. This parameter affects the probability of agents encountering in- and out-group agents. For instance, we can derive the expected probability of an agent from group i encountering their ingroup given their share in population p and parameter φ as:

$$P(i,i) = \varphi \frac{p_i \varphi}{p_i\, \varphi + (1-p_i)(1-\varphi)} + (1-\varphi) \frac{p_i(1-\varphi)}{p_i(1-\varphi) + (1-p_i)\varphi} \tag{37}$$

The first portion of the right side of the equation denotes the probability of agent from group i being in habitat L_i multiplied by the probability of them finding another agent from group i in that same habitat; the second part denotes the probability of agent from group i being in habitat L_j multiplied by the probability of finding another agent from group i in that same habitat. And their expected probability of meeting an outgroup member is:

$$P(i,j) = \varphi \frac{(1-p_i)(1-\varphi)}{p_i\, \varphi + (1-p_i)(1-\varphi)} + (1-\varphi) \frac{(1-p_i)\varphi}{p_i(1-\varphi) + (1-p_i)\varphi} \tag{38}$$

These probabilities are symmetrical for group j. When φ = 0.5, an agent's likelihood of encountering their ingroup is equal to their ingroup's share in the population; when φ = 1, agents will only meet their ingroup, as all agents from group *i* will be in the habitat L_i and all agents from group *j* will be in the habitat L_j. Therefore, the higher the φ, holding the ingroup's share in the total population fixed, the more likely a group member is to encounter their ingroup. Table S2 below shows how parameter φ interacts with *p*, resulting in a range of α values.

φ p	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1**
0.1	0.1	0.11	0.16	0.25	0.45	1
0.2	0.2	0.22	0.29	0.41	0.63	1
0.3	0.3	0.32	0.4	0.52	0.72	1
0.4	0.4	0.42	0.49	0.61	0.78	1
0.5	0.5	0.52	0.58	0.68	0.82	1
0.6	0.6	0.62	0.66	0.74	0.85	1
0.7	0.7	0.71	0.74	0.8	0.88	1
0.8	0.8	0.81	0.82	0.85	0.91	1
0.9	0.9	0.9	0.91	0.92	0.94	1

Table S2. Overview of α values resulting from different combinations of *p* and *φ*

To obtain the likelihood of types of encounters in the population as a whole:

1. The probability of two members of group 1 meeting: $p_1 \cdot P(2,2)$
2. The probability of two members of group 2 meeting: $(1 - p_1) \cdot P(1,1)$
3. The probability of two members from different groups meeting: $p_1 \cdot P(2,1) + (1 - p_1) \cdot P(1,2)$

The three probabilities above add up to 1.

S3.3. Parameter combinations

In our agent-based model, we define trusters' trusting trait so to allow for three truster strategies: unconditional trusting (i.e., always trusting the encountered trustee; trait share within the group denoted with δ_i), unconditional distrusting (i.e., never trusting the encountered trustee; trait share within the group denoted with ε_i), and conditional trusting (i.e., trusting agents with the signal of the same group, distrusting everyone else,; trait share within the group denoted with γ_i). Trustees' signalling trait includes a possibility to signal (trait share within the group denoted 1 – β_i), or not signal (trait share 1 – β_i) group identity. In Table S3, we list all the parameters mentioned in Figure 1 in the main text. As we are interested in the coevolution of traits prescribing sending of signals and traits that help signal recognition, we initialize our populations so that there is no signalling nor conditional trusting in the population (see Table S3, initial population shares).

Parameter	Possible values	Considered values
Constants		
Payoff R_I	$R_I \geq 0$	$R_I = 40$
Payoff R_O	$R_O \geq 0$	$R_O = 30$
Payoff P	$P \geq 0$	$P = 20$
Payoff T	$T \geq 0$	$T = 35$
Payoff S	$S \geq 0$	$S = 5$
Random mutation probability	$m_o \geq 0$	$m_o \in \{0.001, 0.001\}$
Fitness weighting factor (Eq. 35)	$v \geq 0$	$v = 0.9$
N agents	$N \geq 2$	$N = 500$
Trust game rounds	$T > 0$	$T = 20\ 000$
Initial trait shares within each group		
Signalling	$\beta_i \geq 0$	$\beta_1 = \beta_2 = 0$
Conditional trusting	$\gamma_i \geq 0$	$\gamma_1 = \gamma_2 = 0$
Unconditional trusting	$\delta_i \geq 0$	$\delta_1 \approx \delta_2 \approx 0.5$
Unconditional distrusting	$\varepsilon_I \geq 0$	$\varepsilon_1 \approx \varepsilon_2 \approx 0.5$
Population-level varying parameters (fixed)		
Share of group 1	$0 < p_1 \leq 1$	$p_1 \in \{0.1, 0.2, 0.3, 0.4\}$ (implies $p_2 \in \{0.9, 0.8\ 0.7, 0.6\}$)
Group assortment	$0.5 < \varphi \leq 1$	$\varphi \in \{0.5, 0.6, 0.7, 0.8, 0.9, 1\}$
Signalling cost	$c \geq 0$	$c \in \{0, 1, 3, 5\}$

Table S3. Overview of model and simulation parameters

Given the results from our theoretical model and the fact that $\alpha^* = 0.43$, we would expect signalling and signal recognition to emerge if the conditions in Table S4 are satisfied:

α =	Signal if $\gamma \leq$				If $\gamma = 1$	Cond. Trust if $\beta \geq$
	$c = 0$	$c = 1$	$c = 3$	$c = 5$	Max c	
0.1	0	0.5	~~1.5~~	~~2.5~~	2	0
0.2	0	0.25	0.75	~~1.25~~	4	0
0.3	0	0.17	0.5	0.83	6	0
0.4	0	0.13	0.38	0.62	8	0
0.6	0	0.08	0.25	0.25	12	0.5
0.7	0	0.07	0.21	0.21	14	0.68
0.8	0	0.06	0.19	0.19	16	0.81
0.9	0	0.06	0.17	0.17	18	0.92

Note: $\alpha^* = 0.43$

Table S4. Overview of values of β under which the signalling equilibrium emerges

We run combinations of the parameters listed above, initializing each combination twenty times, for a total of 4 800 simulation runs. After each twenty rounds, we record the number of members of each group having each individual trait and the payoffs of members with each of the traits.

S4. Additional results

S4.1. Baseline simulations

We run a scenario without any signalling or conditional trusting as a baseline reference for evaluating the effects of signalling norm emergence on the groups in our population. We vary parameters as per Table S3, but run this simulation for 2 500 rounds only, as this gives populations more than sufficient time to reach the equilibrium states. This baseline scenario is in line with Figure S1 and our Eq. 1 and 2 in this Supplementary Material. As suggested by Eq. 2 and derived for the payoffs in our simulations, we see in Panel A of Figure S3 that trusters adopt the unconditional distrust trait if $\alpha \leq 0.43$ (easiest to see by looking at tiles with $\varphi = 0.5$, when there is no ingroup assortment). Panel B in the same figure shows that, if $\alpha > 0.43$, unconditional trust fixates in the population.

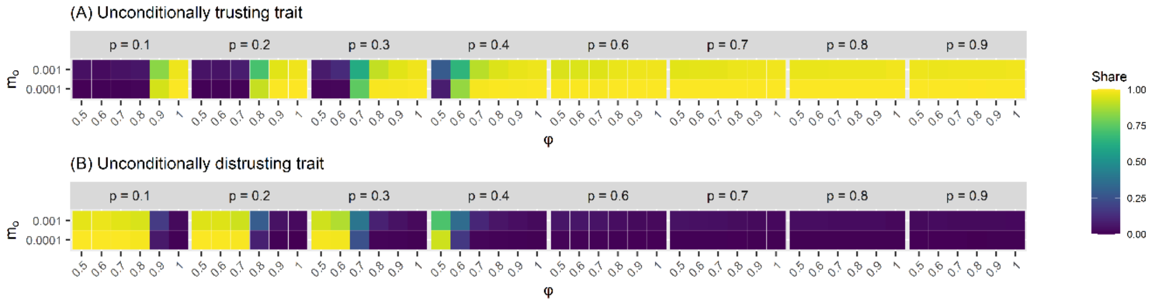


Figure S3. Baseline scenario: Share of agents with a strategy with an: unconditionally trusting trait (A) or unconditionally distrusting trait (B), averaged across the last 100 rounds given the share of ingroup (p) and the assortment parameter (φ) under varying random mutation parameters. We average the results over twenty independent runs of each parameter combination.

S4.2. Signalling and parochial cooperation

Figures S4 and S5 complement Figure 2 in the main text, showing the share of agents adopting strategies other than the one shown in the main text (signalling and conditional trusting). Panel C in Figure S5 corresponds to the panel B in Figure 2 in the main text.

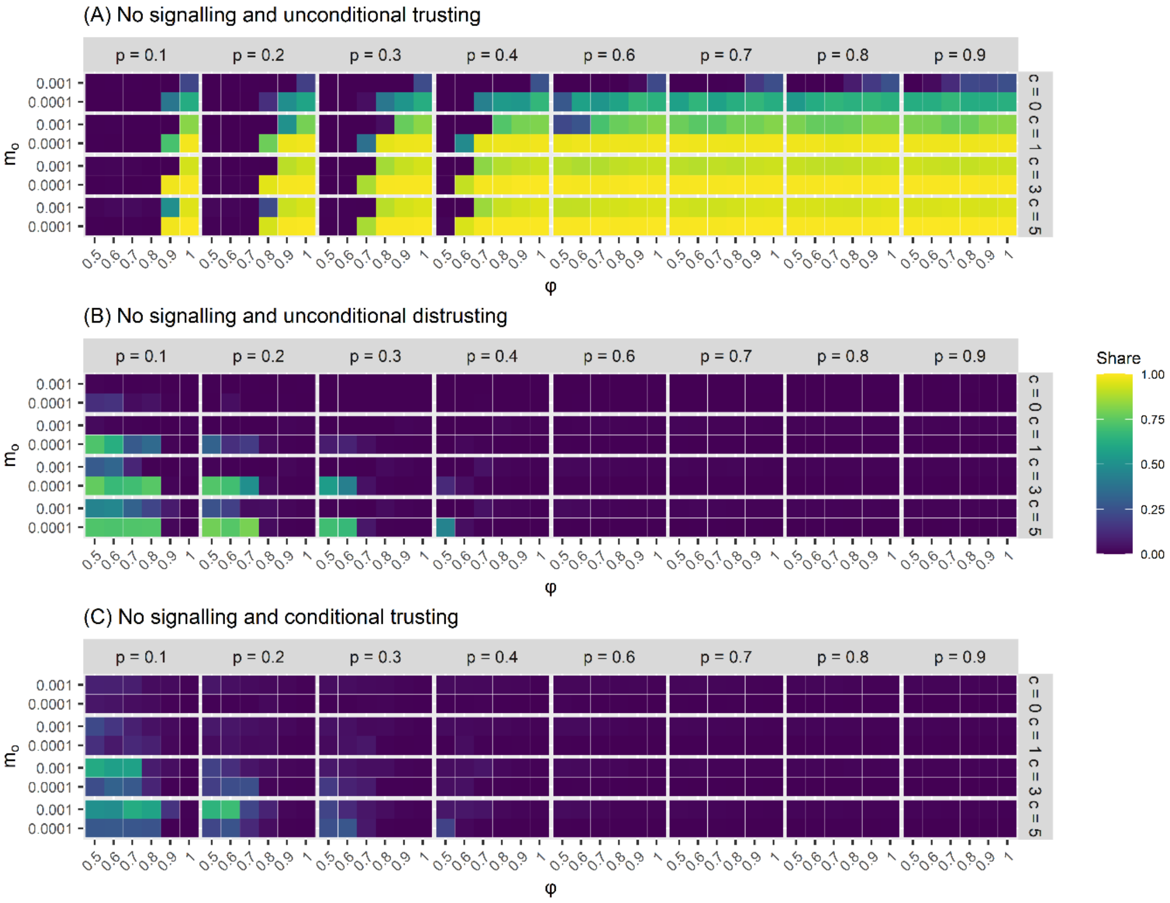


Figure S4. Share of agents with different strategies without signalling trait and various trusting traits averaged across the last 100 rounds given the share of ingroup (p) and the assortment parameter (φ) under varying random mutation parameters. We average the results over twenty independent runs of each parameter combination.

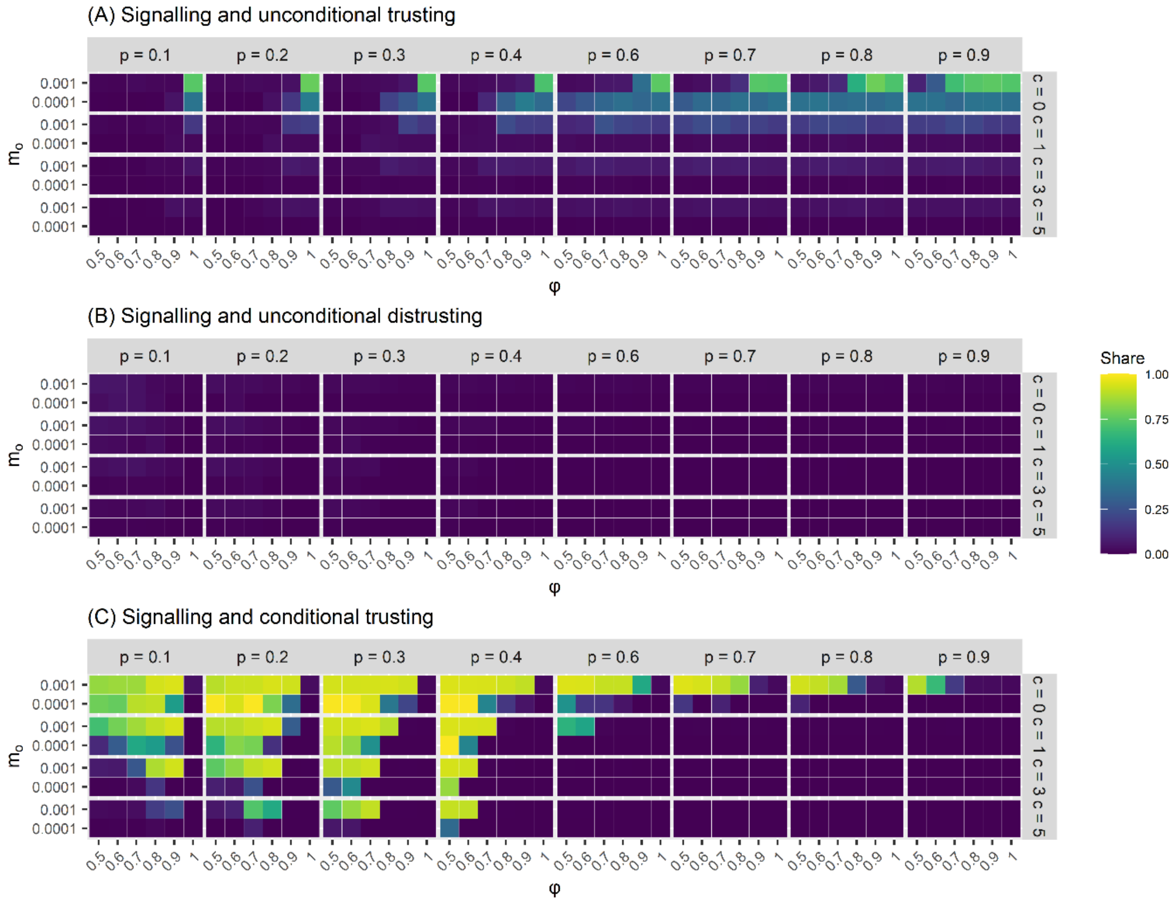


Figure S5. Share of agents with different strategies with a signalling trait and various trusting traits averaged across the last 100 rounds given the share of ingroup (p) and the assortment parameter (φ) under varying random mutation parameters. We average the results over twenty independent runs of each parameter combination.

In Figure S6 below, we show the share of agents with signalling and conditional trusting in the last 100 rounds of the simulation with varying signalling costs, high random mutation rate (m_o = 0.001) as a function of p and φ. Smaller minority groups ($p \leq 0.2$) develop signals at intermediate levels of assortment, once they have a higher chance of running into their own group.

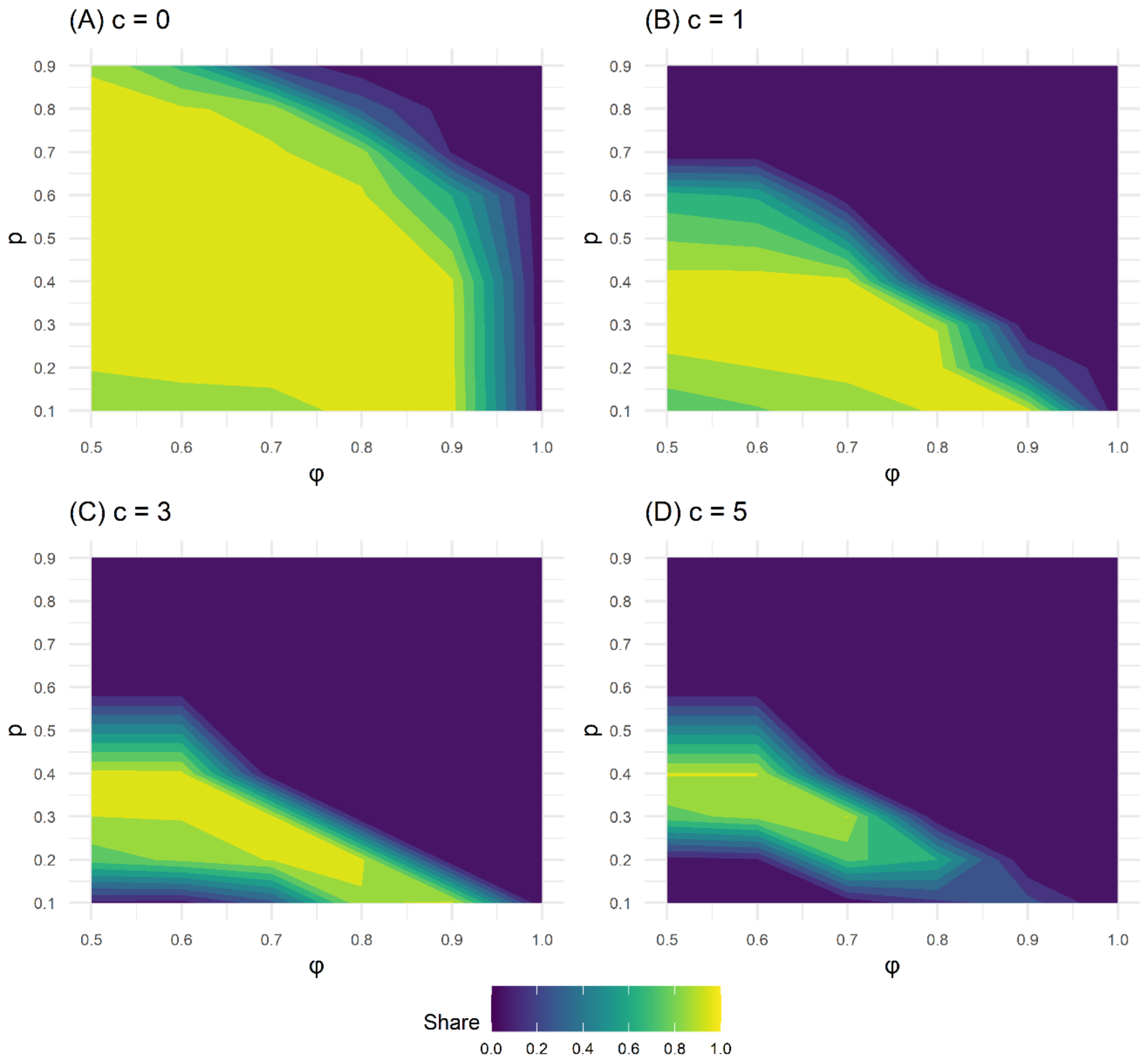
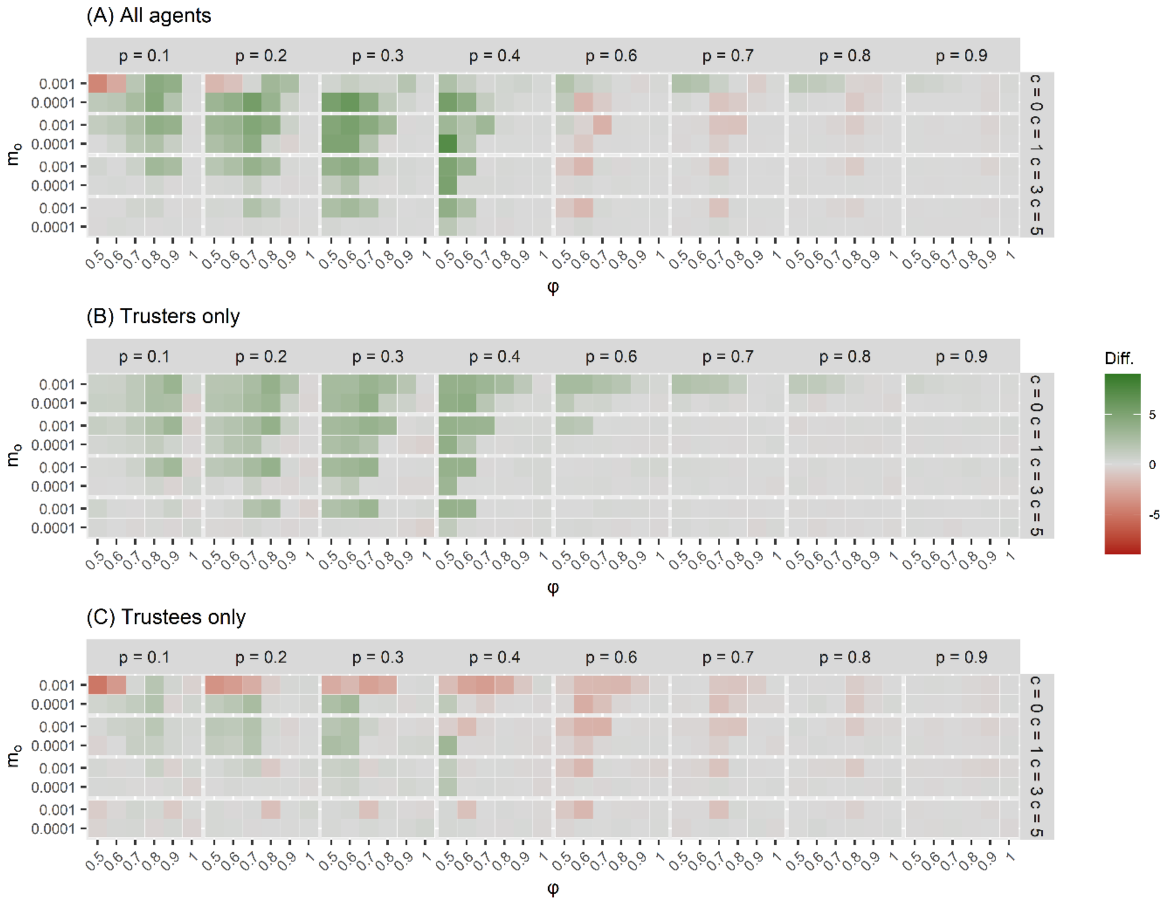


Figure S6. Share of agents with a strategy including signalling and signal recognition (conditional trusting) averaged across the last 100 rounds under different cost regimes (Panels A-D) and under different values of *p* and *φ* parameters. We average the results over twenty independent runs of each parameter combination.

We discuss Panel A of Figure S7 in the main text in the context of differences in average payoffs per agent in the scenario with signalling compared to the baseline scenario. In Panel B we provide differences in average payoffs between these two scenarios for trusters, and in Panel (C) for trustees.



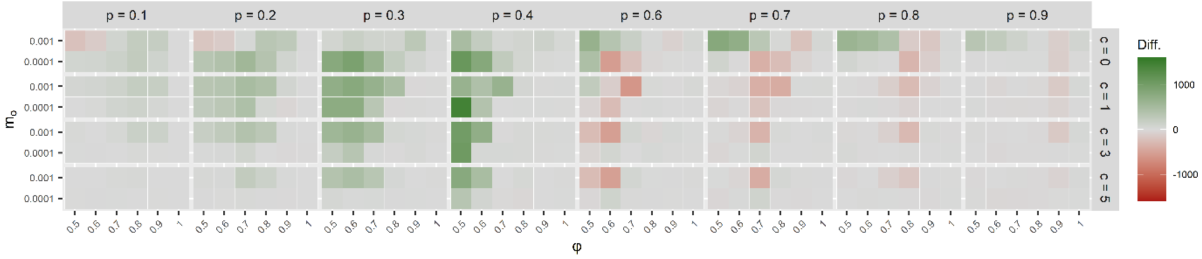
(A) All agents

(B) Trusters only

(C) Trustees only

Figure S7. Difference in average payoffs per different type of group member between (1) the payoffs averaged across the last 100 rounds in a 20 000-round scenario with signalling and (2) the payoffs averaged across the last 100 rounds in a 2 500-round scenario without signalling given the share of ingroup (p) and the assortment parameter (φ) under varying signal costs c and random mutation parameters. We average the results over twenty independent runs of each parameter combination. Tile colouring refers to average payoff differences between regimes with and without signalling. Note that, because groups are interdependent (e.g., groups with p = 0.1 and 0.9 interact with each other), changes in average payoffs might stem from differences in (signalling) traits between the two regimes developed either within the ingroup, or by the outgroup.

In Figure S8 we additionally provide differences in total payoffs obtained by groups of different sizes in scenarios with and without signalling.



Tile coloring refers to total payoff differences between regimes with and without signalling. Note that, because total are interdependent (e.g., groups with p = 0.1 and 0.9 interact with each other), changes in total payoffs might stem from differences between the two regimes in (signalling) traits within the group, but also fromdifferences in traits developed by the outgroup.

Figure S8. Difference in total payoffs per group between (1) the payoffs averaged across the last 100 rounds in a 20 000-round scenario with signalling and (2) the payoffs averaged across the last 100 rounds in a 2 500-round scenario without signalling given the share of ingroup (p) and the assortment parameter (φ) under varying signal costs c and random mutation parameters. We average the results over twenty independent runs of each parameter combination.

S4.3. Dynamics of signalling norm emergence

In Figure 3 in the main text we show dynamics of trait evolution if c = 0 and c = 3 and m_o = 0.001. In Figures S9-S12 below we provide similar plots of trait evolution (including unconditional trusting and distrusting traits) with 95% confidence intervals and for all signal costs and m_o = 0.001.

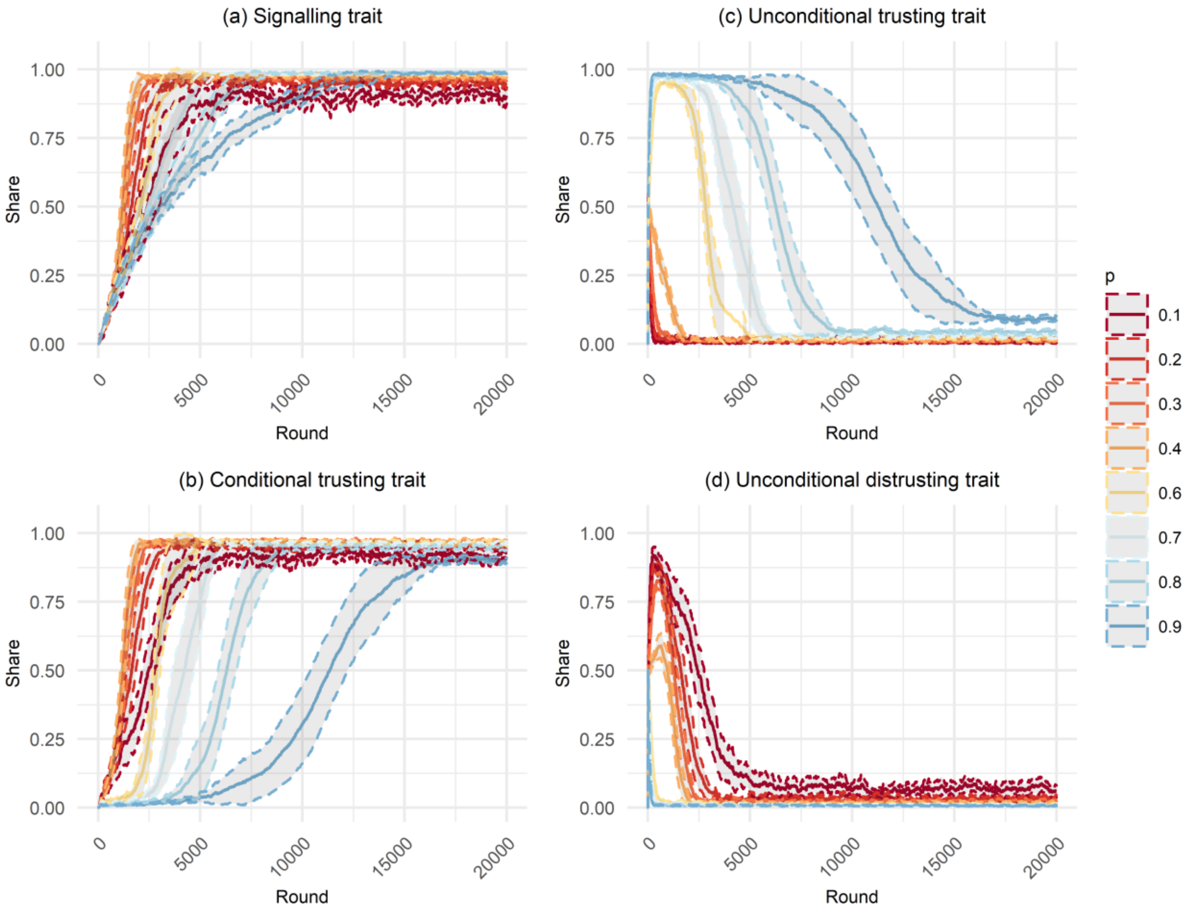


Figure S9. Evolution of signalling (a) and different trusting (b, c, d) traits over 20 000 rounds if signals are costless (c = 0), there is no assortment (φ = 0.5) and random mutations are more common (m_o = 0.001). Shares on the Y axis represent share of different traits within each group. We average the results and calculate the 95% CIs over twenty independent runs of each parameter combination.

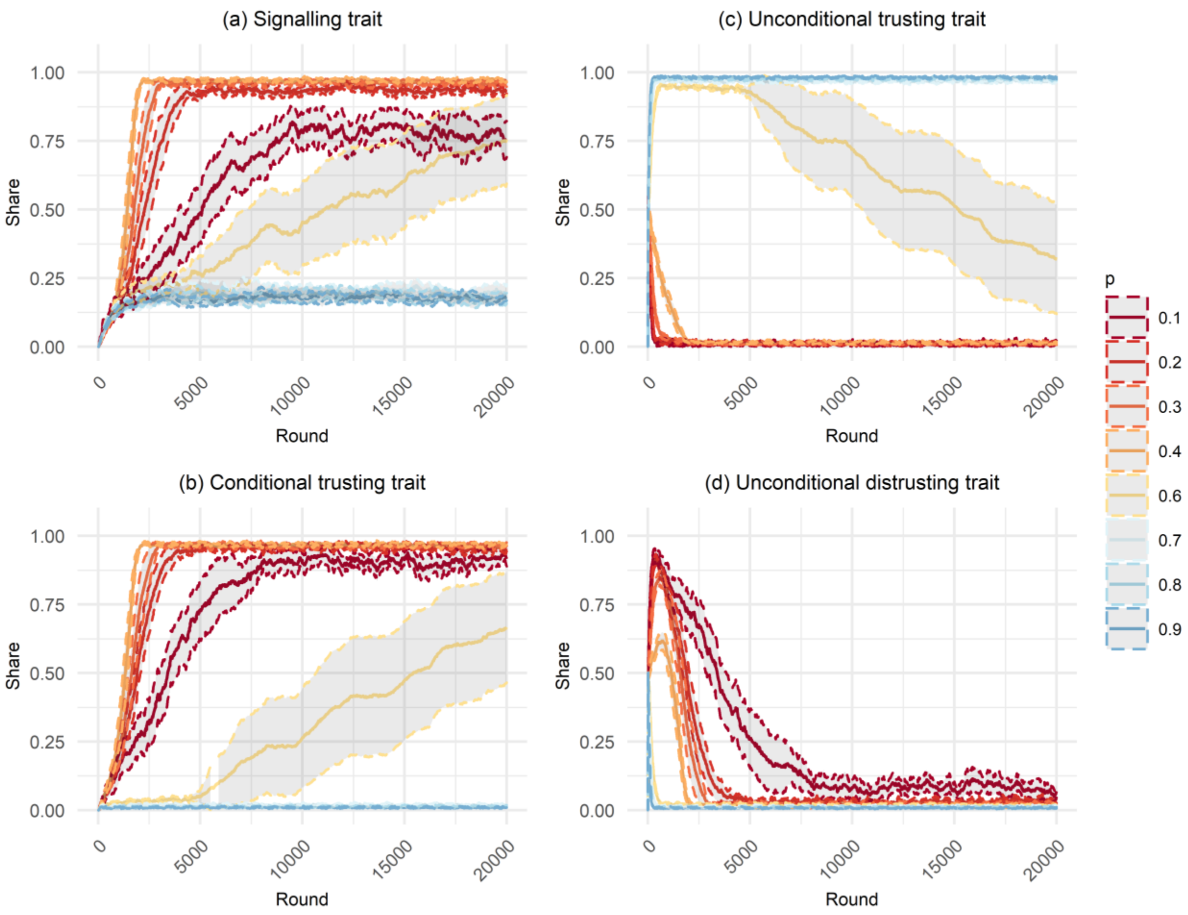


Figure S10. Evolution of signalling (a) and different trusting (b, c, d) traits over 20 000 rounds if signals are costly (c = 1), there is no assortment (φ = 0.5) and random mutations are more common (m_o = 0.001). Shares on the Y axis represent share of different traits within each group. We average the results and calculate the 95% CIs over twenty independent runs of each parameter combination.

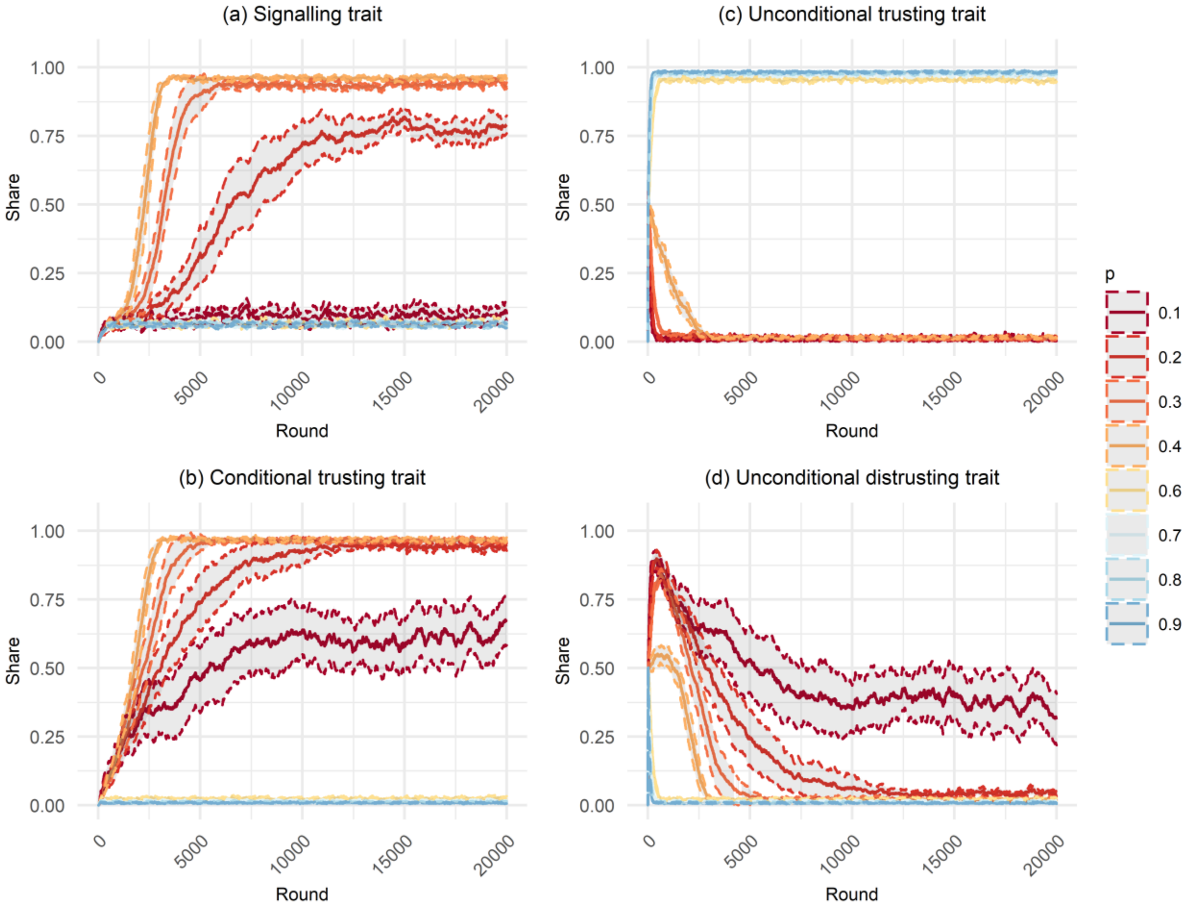


Figure S11. Evolution of signalling (a) and different trusting (b, c, d) traits over 20 000 rounds if signals are costly (c = 3), there is no assortment (φ = 0.5) and random mutations are more common (m_o = 0.001). Shares on the Y axis represent share of different traits within each group. We average the results and calculate the 95% CIs over twenty independent runs of each parameter combination.

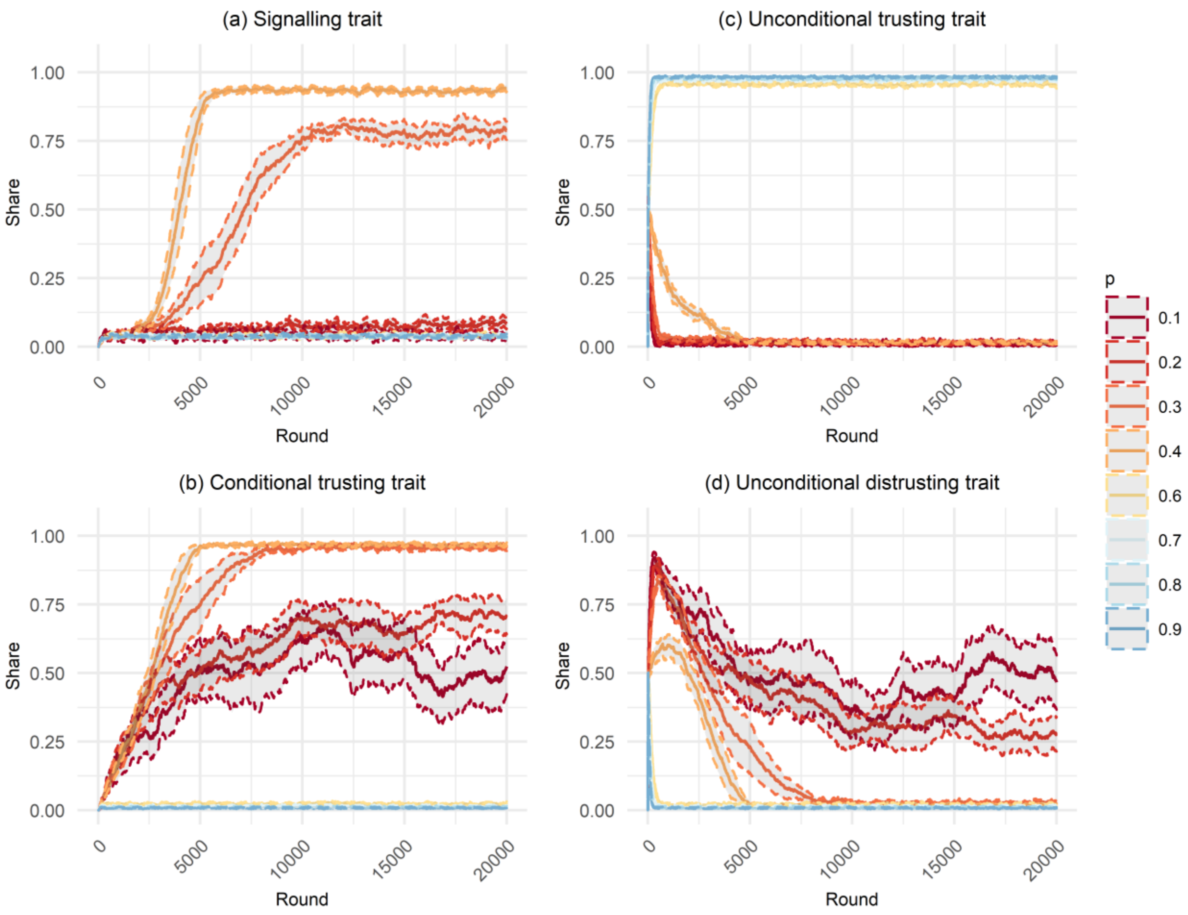


Figure S12. Evolution of signalling (a) and different trusting (b, c, d) traits over 20 000 rounds if signals are costly (c = 5), there is no assortment (φ = 0.5) and random mutations are more common (m_o = 0.001). Shares on the Y axis represent share of different traits within each group. We average the results and calculate the 95% CIs over twenty independent runs of each parameter combination.

S4.4. Dynamics of individual simulation runs

In Figures S13-S16 below, we show two examples of the evolution of different traits in individual simulation runs when random mutations are more (m_o = 0.001) or less common (m_o = 0.0001). Comparing Figures S13 and S14 shows that, under the same parameters (when c = 3), signalling norms develop in all runs if random mutations are more common but only in some runs if the mutations are less common. The difference is even starker between Figures S15 and S16 when signalling costs are higher (c = 5).

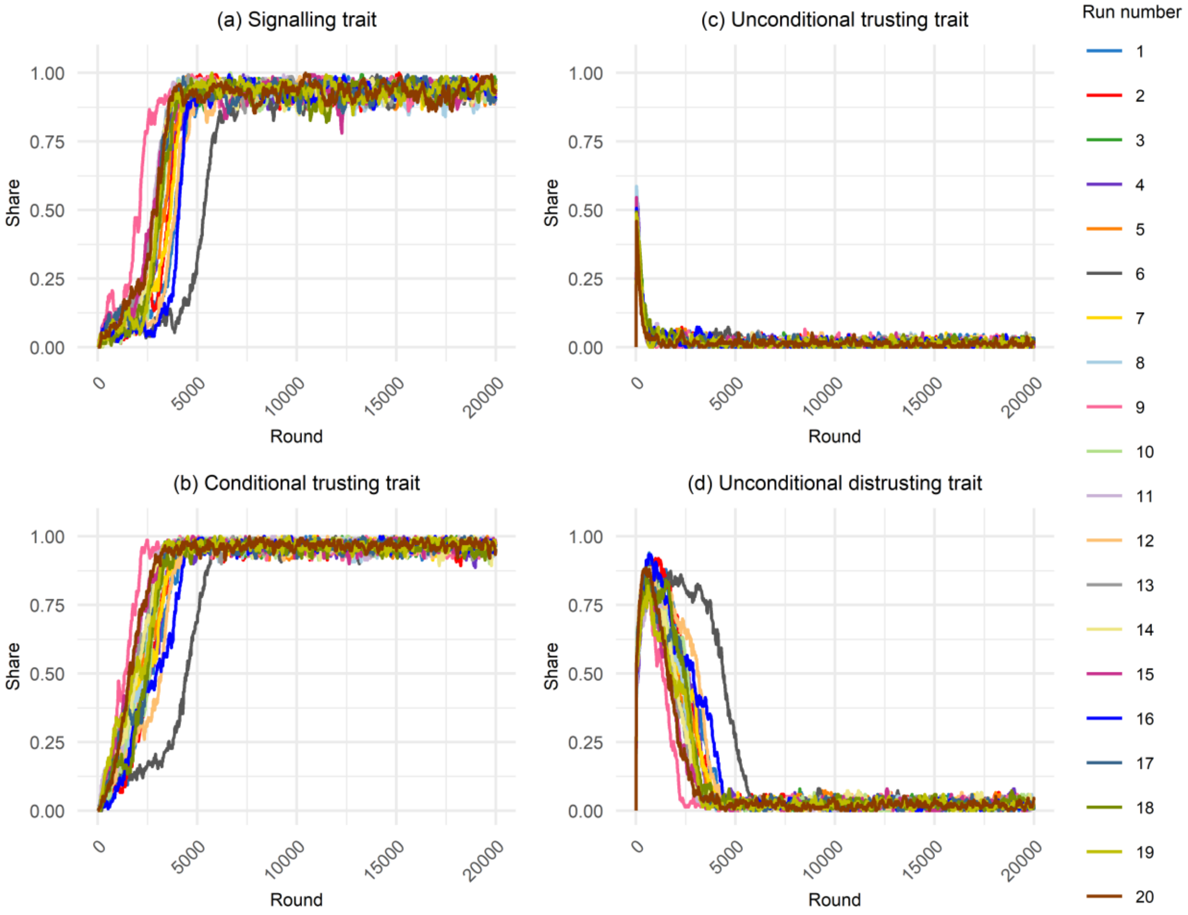


Figure S13. Evolution of signalling (a) and different trusting (b, c, d) traits over 20 000 rounds if signals are costly (c = 3), the group makes up 30% of the population (p = 0.3), there is no assortment (φ = 0.5) and random mutations are more common (m_o = 0.001). Individual lines represent individual independent simulation runs. Shares on the Y axis represent share of different traits within each group.

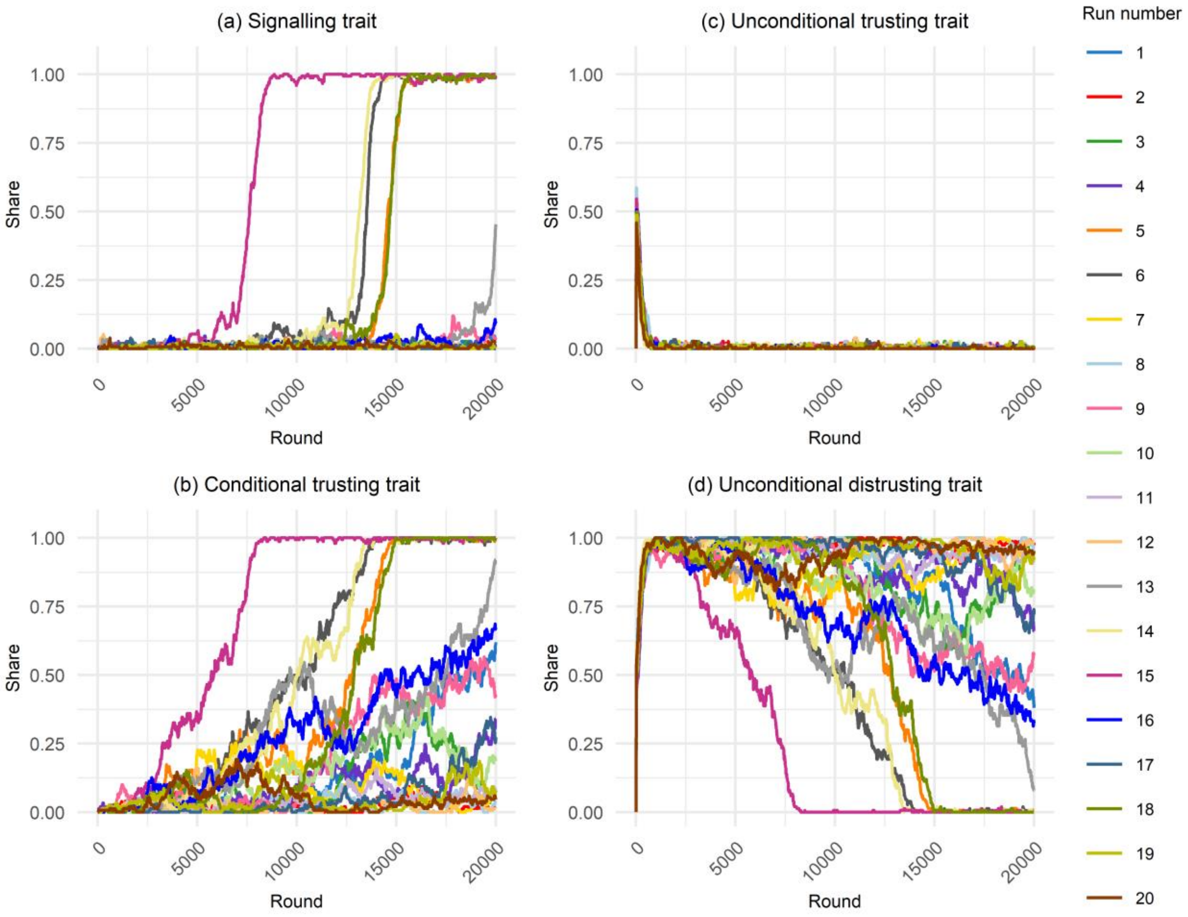


Figure S14. Evolution of signalling (a) and different trusting (b, c, d) traits over 20 000 rounds if signals are costly (c = 3), the group makes up 30% of the population (p = 0.3), there is no assortment (φ = 0.5) and random mutations are less common (m_o = 0.0001). Individual lines represent individual independent simulation runs. Shares on the Y axis represent share of different traits within each group.

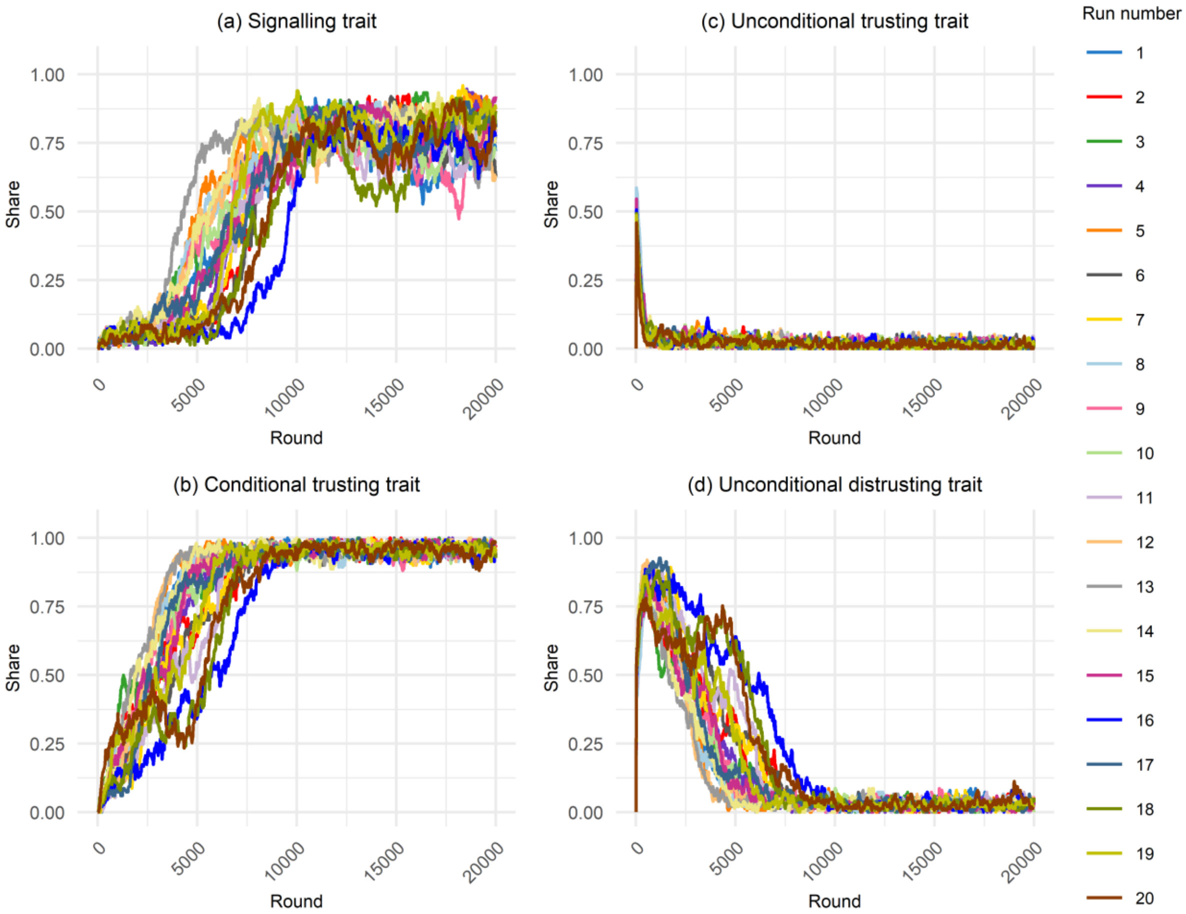


Figure S15. Evolution of signalling (a) and different trusting (b, c, d) traits over 20 000 rounds if signals are costly (c = 5), the group makes up 30% of the population (p = 0.3), there is no assortment (φ = 0.5) and random mutations are more common (m_o = 0.001). Individual lines represent individual independent simulation runs. Shares on the Y axis represent share of different traits within each group.

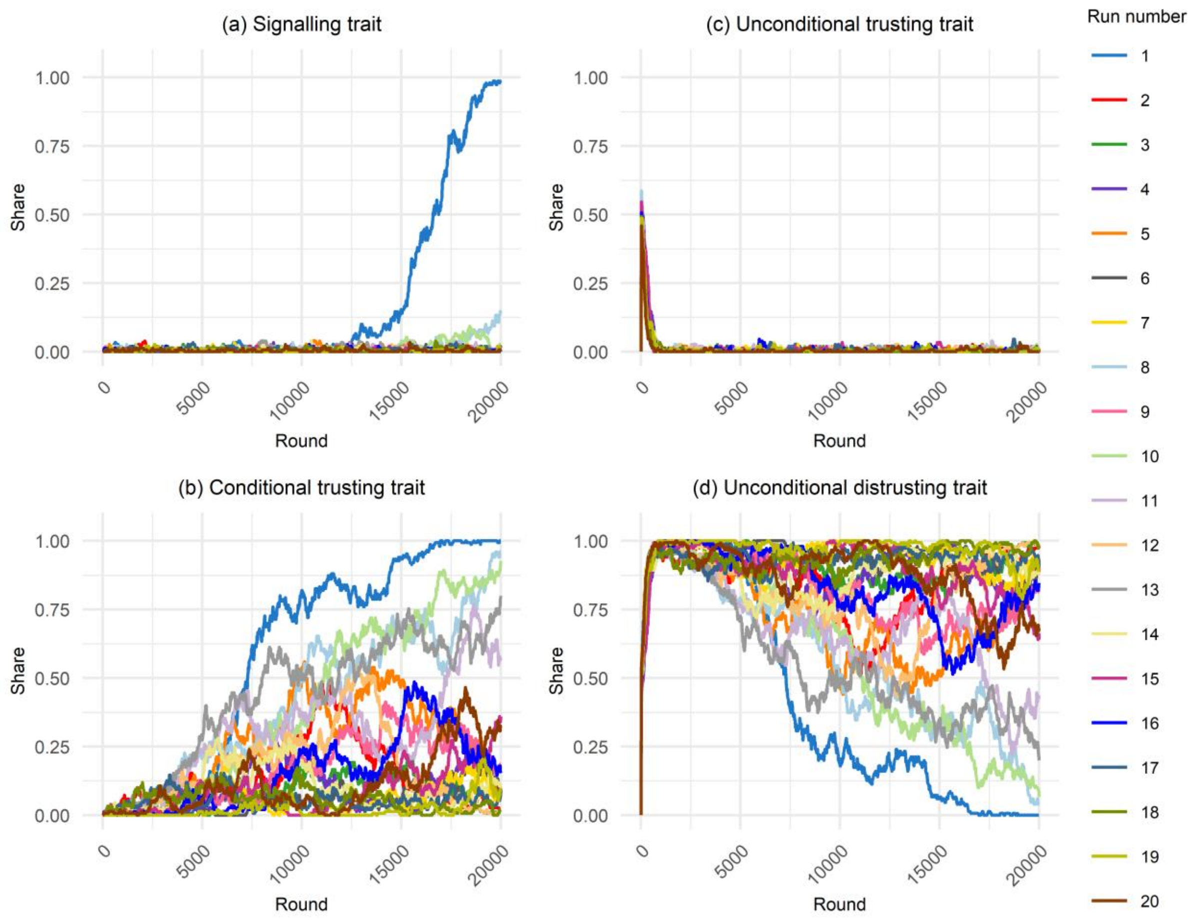


Figure S16. Evolution of signalling (a) and different trusting (b, c, d) traits over 20 000 rounds if signals are costly (c = 5), the group makes up 30% of the population (p = 0.3), there is no assortment (φ = 0.5) and random mutations are less common (m_o = 0.0001). Individual lines represent individual independent simulation runs. Shares on the Y axis represent share of different traits within each group.

S4.5. Evolution under different combinations of p and φ resulting in the same α

Figures S17 and S18 plot each of our twenty independent simulation runs separately, showing that groups with the same α parameter resulting from different combinations of p and φ show different dynamics of signalling and conditional trait evolution. In Figure S17, under the higher random mutation setting (m_o = 0.001), in smaller groups (lower p), signalling-related traits take longer to evolve and there is more variability between the runs. Figure S18 shows how, under low probability of random mutations (m_o = 0.0001), larger groups with lower φ still develop signalling in most runs, while smaller groups with higher φ fail to develop a signalling norm in any of the 20 runs.

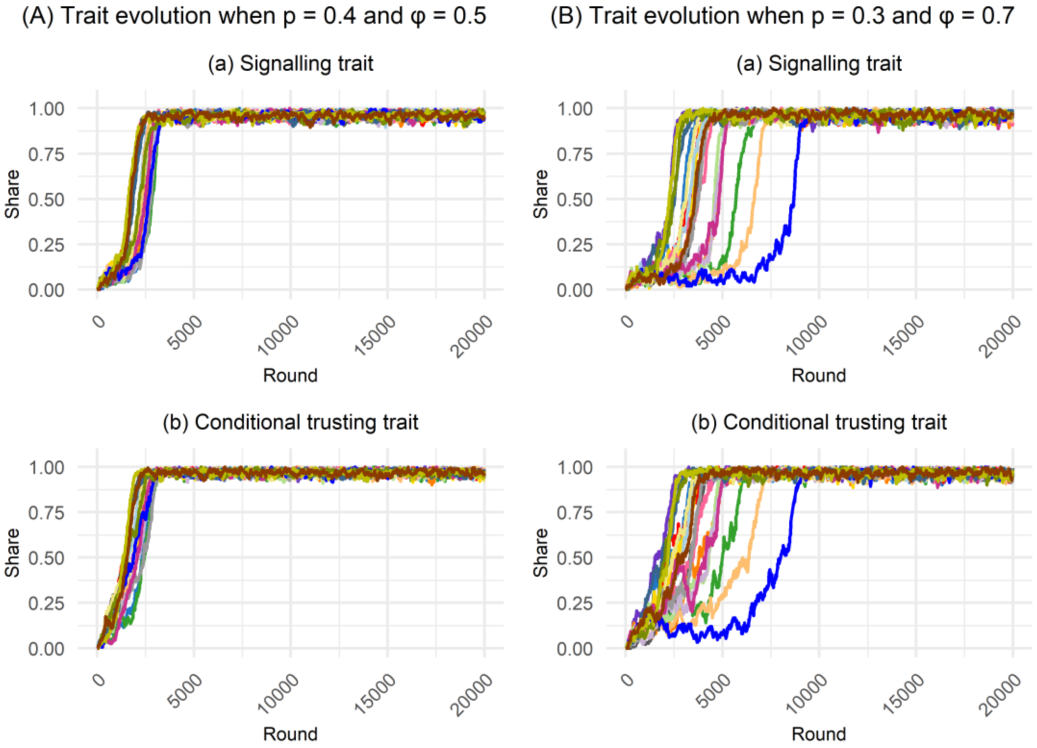


Figure S17. Evolution of signalling (a) and conditional trusting (b) traits over 20 000 rounds if signals are costly (c = 3), and random mutations are more common (m_o = 0.001) under two different combinations of the p and φ parameters, both of which result in α = 0.4. Individual lines represent individual independent simulation runs. Shares on the Y axis represent share of different traits within each group.

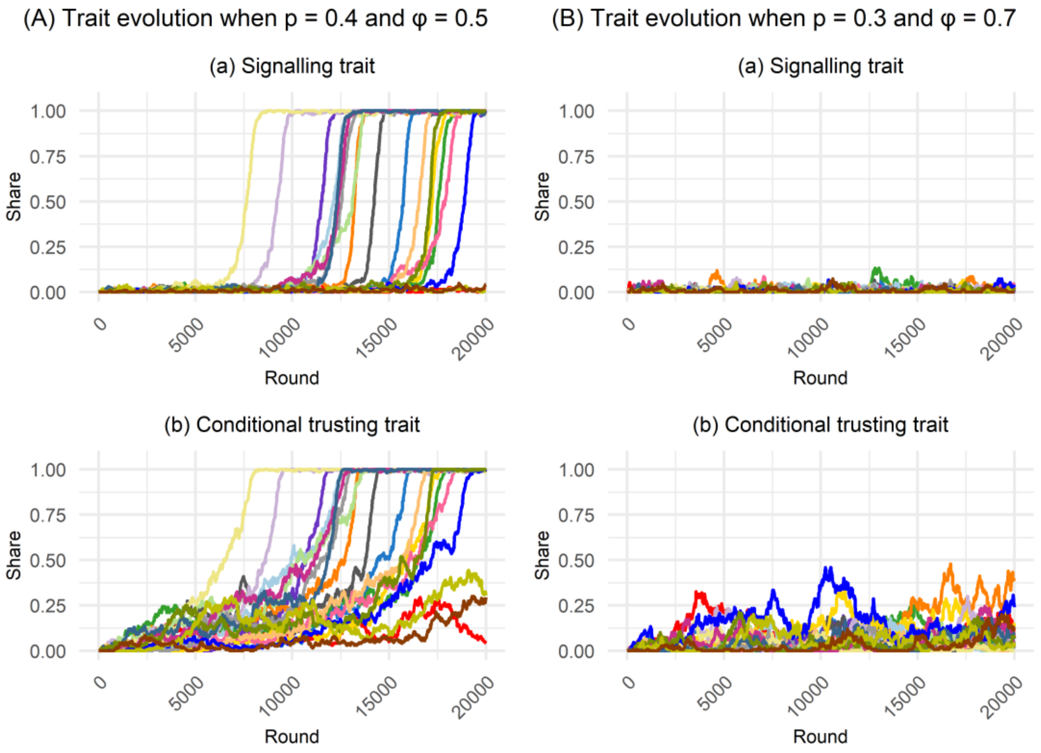


Figure S18. Evolution of signalling (a) and conditional trusting (b) traits over 20 000 rounds if signals are costly (c = 3), and random mutations are less common (m_o = 0.0001) under two different combinations of the p and φ parameters, both of which result in α = 0.4. Individual lines represent individual independent simulation runs. Shares on the Y axis represent share of different traits within each group.

S4.6. Start from a signalling equilibrium

Figure 2A in the main text shows expectations regarding the emergence of the signalling norm based on our analytical results. Our results in Figure 2B trace the emergence of the norm in groups that start in a non-signalling state, and show that not all groups manage to shift into a signalling equilibrium. In Figure S19 below, we show the outcomes of a simulation where both groups were initiated to be in a signalling norm equilibrium with all trustees signalling ($\beta_1 = \beta_2 = 1$) and all trusters conditionally trusting ($\gamma_1 = \gamma_2 = 1$). We run these simulations for 20 000 rounds, with mutation rates as per the Y axis. Figure S19 shows a clear match to our theoretical expectations – signalling and conditional trusting remain largely predominant in groups that were initially in a signalling equilibrium. The exceptions are the small minority groups for which higher signalling costs become unbearable (see the red tiles in Figure 2A in main text) and groups where α approaches 1. As α becomes this high, even small deviations from full signalling and conditional trusting (especially introduced in simulations with higher rates of random mutations) can shift the group into a non-signalling equilibrium (also recall Tables S1 and S4). Since α is sufficiently high in these groups (i.e., the ingroup is almost always encountered), even a small decrease in the share of conditionally trusting trusters deems trustees unwilling to bear the costs of signalling if they encounter their ingroup with full certainty.

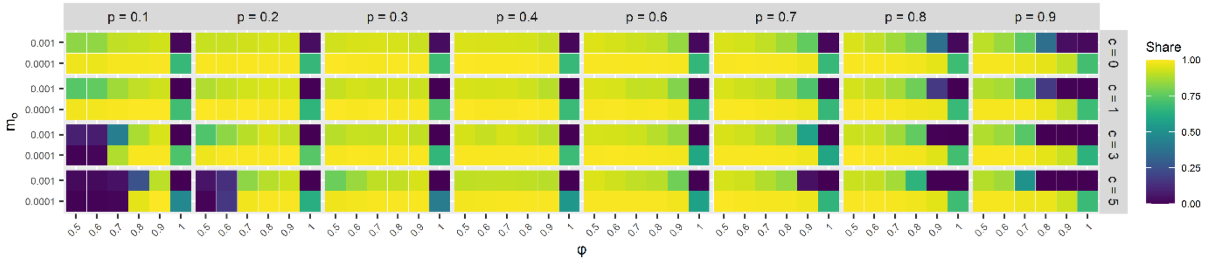


Figure S19. Simulations from a signalling start. The share of agents with a strategy including both signalling and trusting conditional on signals averaged across the last 100 rounds given the share of ingroup (p) and the assortment parameter (φ) under varying signal costs c and random mutation parameters. We average the results over twenty independent runs for each parameter combination.

Combining the results from the simulations starting in a signalling (Figure S19) and non-signalling equilibria (Figure 2B) allows us to draw conclusions on the sizes of the basins of attraction of these two equilibria. It is apparent that, in line with the expectations from our main analysis, the basin of attraction of the signalling equilibrium is larger for groups with $\alpha \leq \alpha^*$ than for groups with $\alpha > \alpha^*$ since even a very small share of signalling and conditional trusting traits in the group (introduced through random mutations) can shift the whole group towards the signalling equilibrium; yet, the opposite is not true for groups with $\alpha > \alpha^*$. As α increases towards

1, the basin of attraction of the signalling equilibrium shrinks, and the basin of attraction of the non-signalling equilibrium becomes larger.

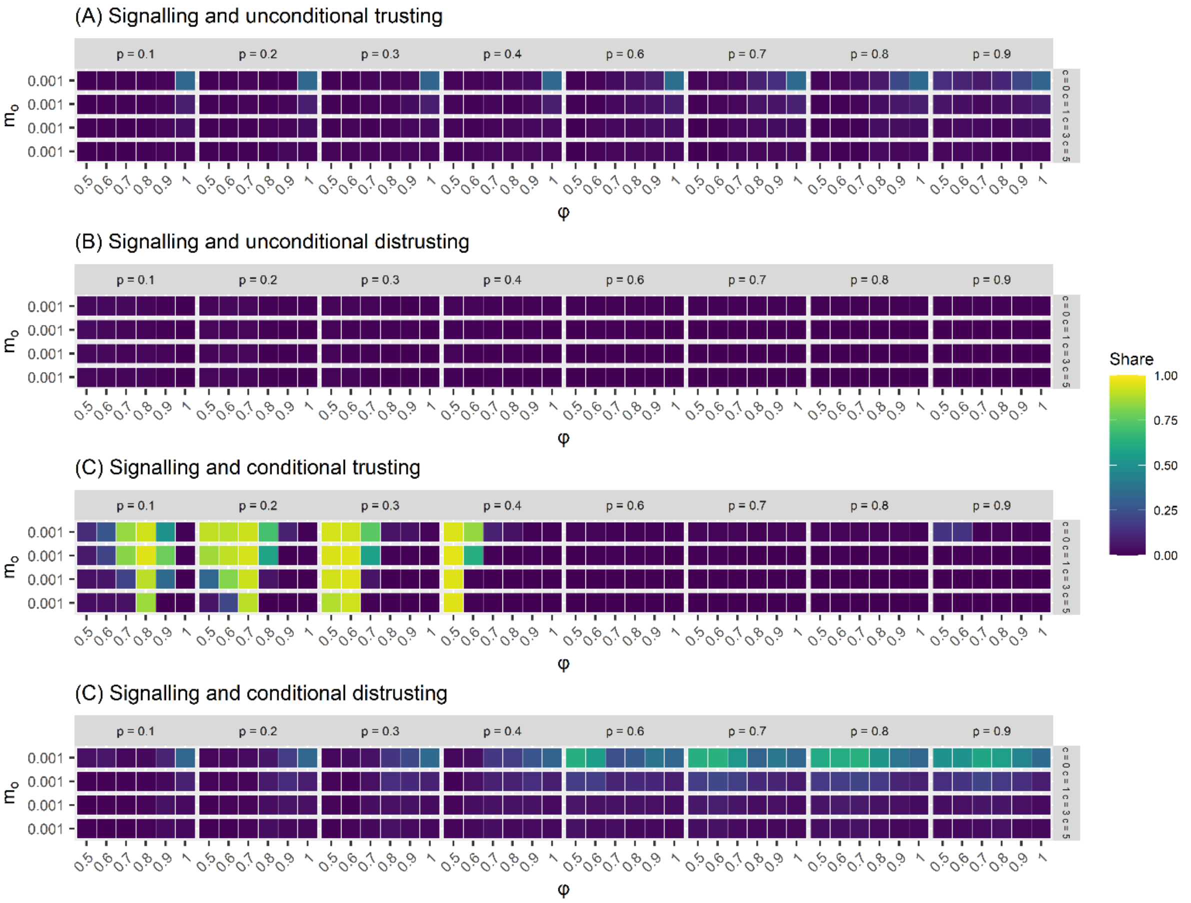
S4.7. Distrusting outgroup signals – simulations

We run additional simulations of the model described in Section 4 in the main text and in Section S2.2 of this Supplementary Material. To account for the different signal cost criteria compared to the main model, we adjust the trust game payoffs as per Table S5 below. We choose these parameters as they maintain a similar value of α^* as in the original setup (0.42 compared to 0.43), but also help satisfy the signalling cost criterium in Eq. 20. All the other parameters remain the same. Due to the computationally demanding nature of the task, we only survey the outcomes of simulations with the higher random mutation probability (m_o = 0.001).

Parameter	Possible values	Considered values
Constants		
Payoff R_I	$R_I \geq 0$	$R_I = 65$
Payoff R_O	$R_O \geq 0$	$R_O = 33$
Payoff P	$P \geq 0$	$P = 30$
Payoff T	$T \geq 0$	$T = 35$
Payoff S	$S \geq 0$	$S = 5$

Table S5. Overview of payoff parameters in additional simulations

As expected given our analyses, Panel C in Figure S20 shows that the range of groups that develop signalling is smaller compared to our baseline model. Groups with p = 0.1 do not adopt signalling even at low costs ($c \leq 1$), as their benefits of parochial cooperation do not offset the losses they incur upon encountering the conditionally distrusting outgroup. Further, we see lower emergence of signalling norms in minority groups even at moderate increases in assortment. If signals are costless, groups with p > 0.4 might still adopt them, but rather than fully parochially cooperating, these majority groups tend to adopt a high share of conditional distrusting in response to minority group's attempts at signalling, which prevents the full emergence of signalling norms.



(A) Signalling and unconditional trusting

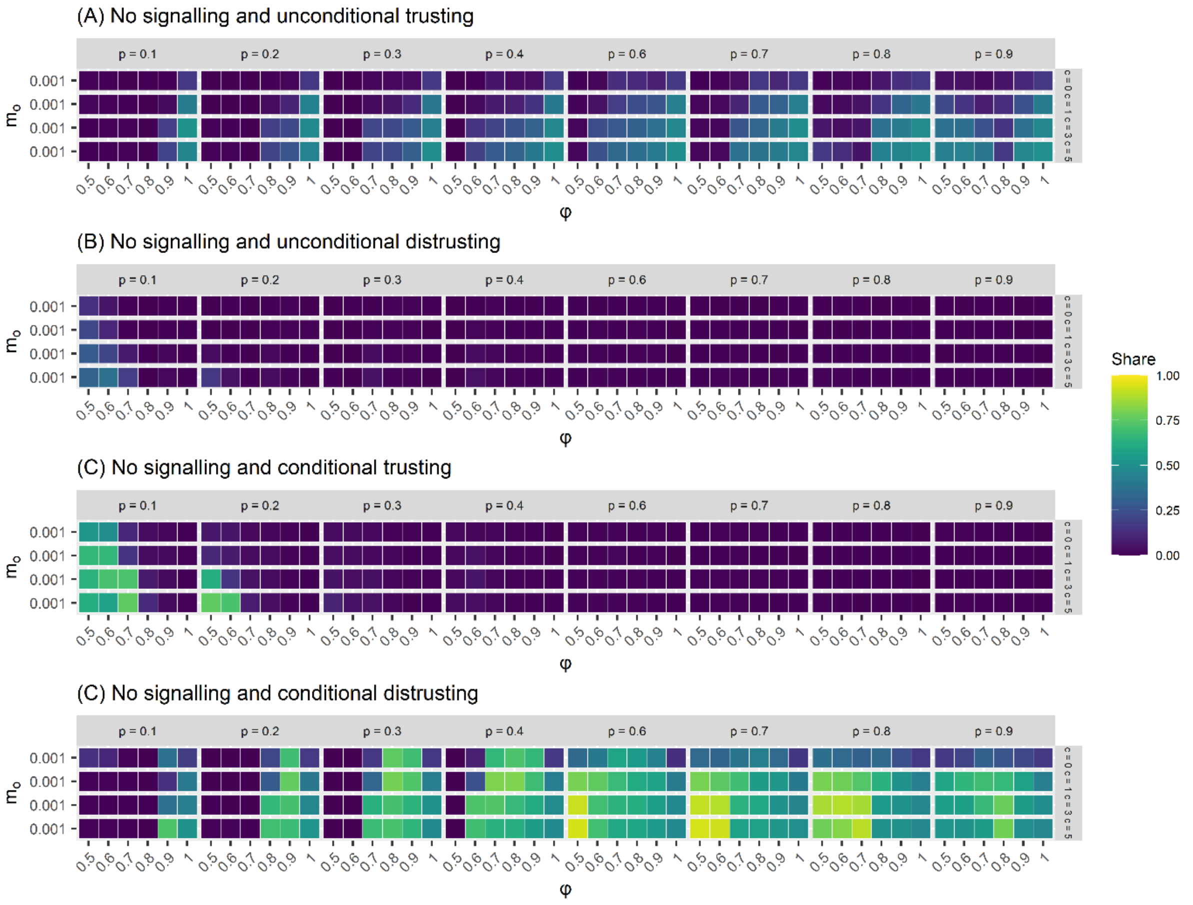
(B) Signalling and unconditional distrusting

(C) Signalling and conditional trusting

(C) Signalling and conditional distrusting

Figure S20. The share of agents with different strategies with a signalling trait and various trusting traits averaged across the last 100 rounds given the share of ingroup (p) and the assortment parameter (φ) under varying random mutation parameters. We average the results over twenty independent runs of each parameter combination.

Figure S21 further helps us understand which strategies are adopted by different groups. We see that, unlike in our main model, all groups that do not develop a signalling norm adopt a large share of conditional distrusting – even if they do not face a signalling outgroup (e.g., see Panel C for a group with p = 0.9 facing a non-signalling and mostly conditionally trusting outgroup with p = 0.1 at φ = 0.5). These results suggest that, as random mutations introduce outgroup signalling, groups that would otherwise remain unconditionally trusting become unconditionally distrusting in order to avoid infrequent, but present, abuse of their trust. As a result, groups become unconditionally distrusting against the outgroup – even if this outgroup is not signalling. Even small bouts of minority signalling – without the full establishing of the norms – can bring upon wide-spread discrimination by the outgroup.


(A) No signalling and unconditional trusting

(B) No signalling and unconditional distrusting

(C) No signalling and conditional trusting

(C) No signalling and conditional distrusting

Figure S21. The share of agents with different strategies with no signalling trait and various trusting traits averaged across the last 100 rounds given the share of ingroup (p) and the assortment parameter (φ) under varying random mutation parameters. We average the results over twenty independent runs of each parameter combination.

S4.8. Noise in truster recognition

Using Eq. 28, Eq. 32, and Eq. 34 we can calculate the critical values of ζ for which signalling equilibrium holds in groups with different α. Table S6 shows that, given the payoffs in our main simulations, signalling equilibrium as per Theorem 9 exists for groups with all α values we survey in while $0 \leq \zeta \leq 0.5$.

α	**General ζ*** (Eq. 28)	**Group-specific ζ** (Eq. 32 if $\alpha \leq 0.43$ and Eq. 34 if $\alpha > 0.43$)
0.1	0.61	0.57
0.2	0.62	0.57
0.3	0.64	0.57
0.4	1	0.57
0.6	0.55	0.6
0.7	0.56	0.6
0.8	0.57	0.6
0.9	0.57	0.6

Table S6. Overview of threshold noise parameters ζ. If the extent to which trustees incorrectly perceive trusters' group identities upon being given trusts exceeds the values in this table, signalling and signal recognition are not a part of the equilibrium we define in Theorem 9.

Do note that there are payoff combinations where there could be ζ values that do not satisfy the conditions in Eq. 28, 32, and 34,. For instance, keeping all other payoffs in our setup constant, setting S = 15 would result in ζ^* = 0.23.

In Figure S22 below, we show that the outcomes with regard to the emergence of signalling in our agent-based simulations are robust to increases in the rate of trustee misperception of trusters ingroup identity, just as we expected given our analytical results in Section S2.3. We see that, in all cases where signalling emerges fully without when there is no noise, increasing the share of cases in which trustees perceive truster's group identity wrong (honouring outgroup trust and dishonouring ingroup trust) up to 40 percent has no bearing on the final shares of signalling norms. In some cases, where emergence is incomplete even without any noise (e.g., at higher φ values when $p \in \{0.1, 0.2\}$ and c = 5, at lower φ values when p = 0.6 and c = 1, and at $p \in \{0.6, 0.7, 0.8, 0.9\}$ and c = 0), we see some variability in signalling norm emergence given different error rates. Finally, we do see that, for some majority groups, introducing high rates of noise in truster recognition can help sustain low to moderate rates of emergence of signalling under low signalling costs (e.g., p = 0.7 and c = 1).

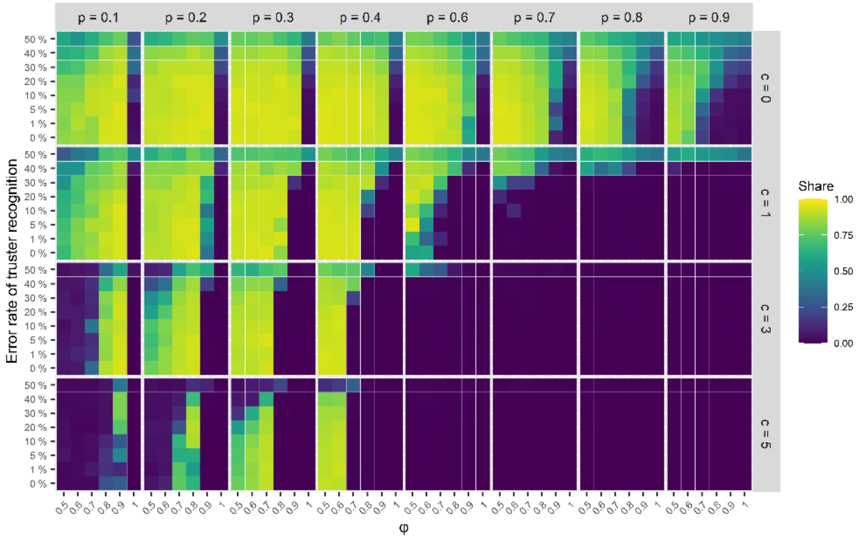


Figure S22. The share of agents with a strategy including both signalling and trusting conditional on signals averaged across the last 100 rounds given the share of ingroup (p) and the assortment parameter (φ) under varying signal costs c and random mutation parameters. We average the results over ten independent runs for each parameter combination, with random mutation rate set to $m_0 = 0.001$.

As ζ approaches ζ^* as per Eq. 28 for $p \geq 0.6$, these groups approach the scenario where trusters prefer to unconditionally distrust, instead of unconditionally trusting, all trustees. Note that this resembles the scenario when $\alpha \leq \alpha^*$ as per our main analyses. The fact that trusters are better off distrusting all trustees makes these groups, with $\alpha > \alpha^*$, but $\zeta \approx \zeta^*$ susceptible to developing a signalling norm, not unlike groups with $\alpha \leq \alpha^*$ when $\zeta = 0$ or $\alpha \leq \alpha^*$ and $\zeta \leq \zeta^*$.

S4.9. Agent-based simulation robustness checks

We check how robust our agent-based simulation outcomes are to changes in the values of some main parameters. First, we vary the number of agents in the population so that $N \in \{250, 500, 1000\}$, and the rate of random mutations so that it includes an additional value $m_o \in \{0.0001, 0.001, 0.002\}$. Due to the computationally demanding nature of this task, we run these simulations for 15 000 steps (instead of 20 000 in the main simulations) and initiate 10 independent simulation runs. In Figure S23 we show that changes to the number of agents in our population do not alter our main conclusions at higher mutation rates. It is only when it comes to parameter combinations where the emergence is not full (e.g., p = 0.3, φ = 0.5, m_o = 0.001) that we see more variability between simulations with different numbers of agents.

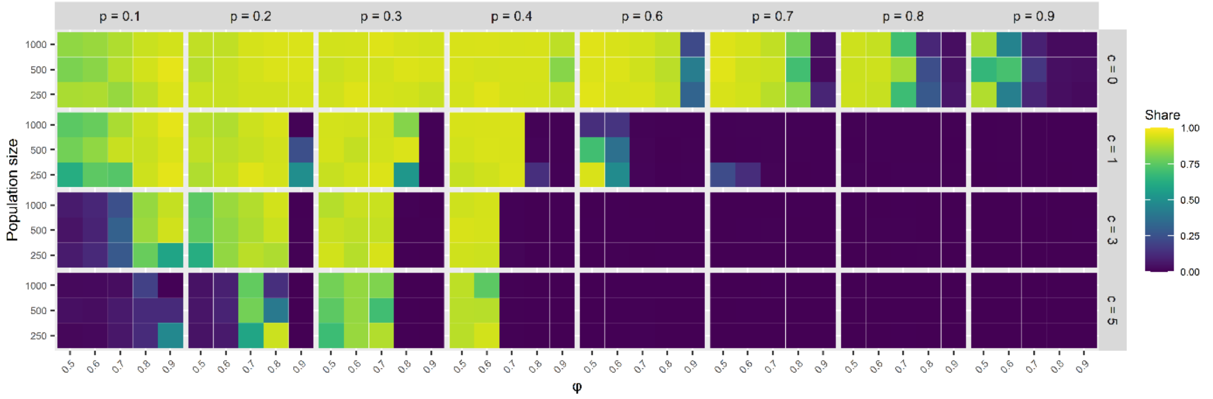


Figure S23. The share of agents with a strategy including both signalling and trusting conditional on signals averaged across the last 100 rounds (out of 15 000) given different population sizes (N) and the assortment parameter (φ) under varying signal costs c. We average the results over ten independent runs for each parameter combination, with random mutation rate set to m_o = 0.001.

Figure S24 shows that our results are mostly robust to the increase in mutation rates. Once again, we see more variability in scenarios where the signalling norm evolution is not as straightforward in the main simulations either. Here, either higher or lower rates of mutation can lead to somewhat different outcomes. Overall, however, we can conclude that increasing the mutation rate does not lead to qualitatively different results; while our main conclusion about the necessity of sufficiently high mutation rates still holds.

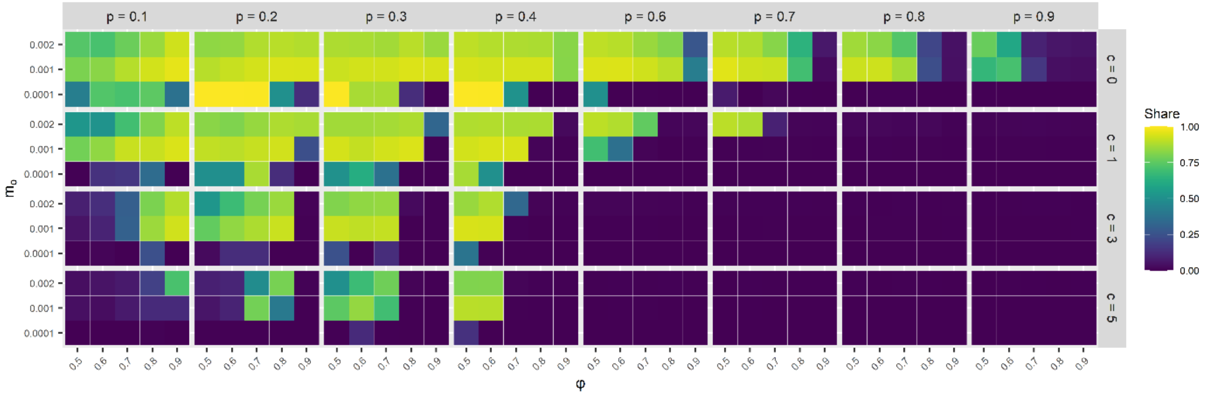


Figure S24. The share of agents with a strategy including both signalling and trusting conditional on signals averaged across the last 100 rounds (out of 15 000) given different mutation rates (m_o) and the assortment parameter (φ) under varying signal costs c. We average the results over ten independent runs for each parameter combination, with number of agents N = 500.

Finally, in Figure S25, we evaluate simulation outcomes given different mutation rates and numbers of agents. Once again, at higher mutation rates (when $m_o \in \{0.001, 0.002\}$), we see that changes in the number of agents in the simulation lead to comparable outcomes in most scenarios. We do see that, in some boundary cases (e.g., p = 0.6), signalling norms emerge more readily in smaller populations (N = 250). At lower mutation rates (m_o = 0.0001), there is a larger variability of outcomes.

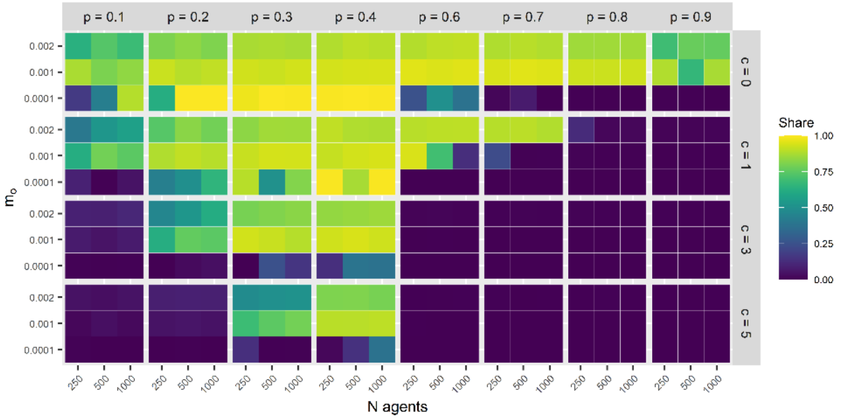


Figure S25. The share of agents with a strategy including both signalling and trusting conditional on signals averaged across the last 100 rounds (out of 15 000) given different mutation rates (m_o) and population sizes (N) under varying signal costs c. We average the results over ten independent runs for each parameter combination, with assortment parameter φ = 0.5.

References

1. Przepiorka W, Diekmann A. 2021 Parochial cooperation and the emergence of signalling norms. *Philos. Trans. R. Soc. B Biol. Sci.* **376**. (doi:10.1098/rstb.2020.0294)

2. Przepiorka W, Berger J. 2017 Signaling Theory Evolving: Signals and Signs of Trustworthiness in Social Exchange. In *Social dilemmas, institutions, and the evolution of cooperation* (eds B Jann, W Przepiorka), Berlin, Germany and Boston, MA: De Gruyter. (doi:10.1515/9783110472974-018)

3. Wilensky U. 1999 NetLogo.

4. Macy MW, Skvoretz J. 1998 The Evolution of Trust and Cooperation between Strangers: A Computational Model. *Am. Sociol. Rev.* **63**, 638. (doi:10.2307/2657332)

5. Chica M, Chiong R, Adam MTP, Teubner T. 2019 An Evolutionary Game Model with Punishment and Protection to Promote Trust in the Sharing Economy. *Sci. Rep.* **9**, 19789. (doi:10.1038/s41598-019-55384-4)

6. Rendell L *et al.* 2010 Why Copy Others? Insights from the Social Learning Strategies Tournament. *Science* **328**, 208–213. (doi:10.1126/science.1184719)