Polarize, Catalyze, Stabilize

How an active minority of norm internalizers amplifies group selection and punishment

by

Victor Vikram Odouard[1*], Diana Smirnova[2], and Shimon Edelman[3]

[1]Santa Fe Institute

[2]Cornell University, Department of Computer Science

[3]Cornell University, Department of Psychology

* Author to whom correspondence should be addressed: Victor Vikram Odouard, Santa Fe Institute (Cowan Campus) 1399 Hyde Park Road, Santa Fe, New Mexico 87501, U.S.A. E-mail: v047@cornell.edu

Acknowledgments

We would like to thank Michael Macy, Alex Vladimirsky, Michael Holton Price, Tyler Millhouse, and David Krakauer for their thoughts, ideas, and feedback; our anonymous reviewers, whose thorough comments allowed us to greatly improve the paper; and David Krakauer and the Santa Fe Institute, under whose auspices we were able to complete the last painstaking steps of this research.

Funding

This work is supported by a grant to SFI from the Omidyar Network to cover core research in the area of Emergent Political Economy. David Krakauer is the PI on this award. The funders played no role in the design, execution, writing, and submission of this research.

Conflicts of interest

We have no conflicts of interest to disclose.

Data Availability Statement

The datasets generated and analyzed during the current study are available on Google Drive at `https://drive.google.com/drive/folders/191NgPRAGVb0q4hbv9BUPXqfh7lSLAKpv?usp=sharing`. Furthermore, the full codebase for the models, including the code we used to analyze the data, can be found at `https://github.com/victorvikram/norm-internalization-and-coop`.

Abstract

Many mechanisms behind the evolution of cooperation, such as reciprocity, indirect reciprocity, and altruistic punishment, require group knowledge of individual actions. But what keeps people cooperating when no one is looking? Conformist norm internalization, the tendency to abide by the behavior of the majority of the group, even when it is individually harmful, could be the answer. In this paper, we analyze a world where (1) there is group selection and punishment by indirect reciprocity but (2) many actions (half) go unobserved, and therefore unpunished. Can norm internalization fill this 'observation gap' and lead to high levels of cooperation, even when agents may in principle cooperate only when likely to be caught and punished? Specifically, we seek to understand whether adding norm internalization to the strategy space in a public goods game can lead to higher levels of cooperation when both norm internalization and cooperation start out rare. We found the answer to be positive, but, interestingly, not because norm internalizers end up making up a substantial fraction of the population, nor because they cooperate much more than other agent types. Instead, norm internalizers, by polarizing, catalyzing, and stabilizing cooperation, can increase levels of cooperation of other agent types, while only making up a minority of the population themselves.

Significance Statement

What keeps people cooperating when no one is looking? Group selection favoring cooperative groups does not require observers, but often works best when there is some other norm enforcement mechanism to supplement it. But most norm enforcement mechanisms require observers in order to function. Here we examine whether norm internalization could fill this enforcement gap, by acting as an "internal enforcer" of norms. Perhaps surprisingly, we discovered in our simulations that the population of norm internalizers always remained quite small, and norm internalizers didn't necessarily cooperate more than other agent types. Nevertheless, under certain conditions, they were able to spark significantly higher mean levels of cooperation, by causing other agents to cooperate more – partly by sparking bouts of high cooperation after periods of very low cooperation, and partly by helping maintain those bouts of high cooperation for longer.

Keywords

norm internalization · evolution · group selection · cooperation · public goods game · conformist transmission

1 Introduction

Because altruism, as expressed in unconditional cooperation or in self-sacrifice, is by itself not an evolutionarily stable trait, its persistence requires a special explanation (Darwin 2011). Almost all such explanations share the characteristic that they *make altruists more likely to accrue the benefits of altruism from others* (Henrich 2004).

In kin altruism the agent benefits from the altruism of its relatives, who are disproportionately likely to share its altruistic gene (Hamilton 1964; Queller 1992). Reciprocity ensures that the agent benefits from the help of those it helped in the past, either directly (Axelrod 1981; Axelrod and Hamilton 1981; Brown et al. 1982; Trivers 1971), or indirectly, when it benefits from those who act on its reputation as a cooperator (Nowak and Sigmund 2005; Panchanathan and Boyd 2003). Because punishers coerce those around them into cooperating, they, too, are more likely to accrue benefits of altruism (Boyd et al. 2003; Boyd and Richerson 1992). Finally, under group selection altruistic groups are more likely to proliferate, causing the majority of altruists to reside in highly altruistic groups and therefore benefitting collectively from altruism (Bowles et al. 2003; Nowak 2006).

Here we examine another mechanism: conformist norm internalization, in which an agent conforms to the majority behavior of the group (Gintis 2003; Lehmann and Feldman 2008), even if this runs counter to its own (fitness) interest. Because those in cooperative groups thus conform to the cooperative norm, this conformity implies that cooperators are more likely to be in cooperative groups than non-cooperators. But this *consequence* of norm internalization doesn't elucidate why a high population of norm internalization came about in the first place (Lehmann et al. 2008b). In this paper, we examine whether norm internalization can create conditions of higher cooperation *without* making any assumptions about their population: that is, we allow for the possibility that they might die out.

More formally, we ask, assuming the presence of group selection and punishment,

Q1 Can norm internalization and cooperation proliferate when both start out rare, and, if yes,

Q2 Does the presence of norm internalizers (NIs) increase cooperation levels over and above the effects of group selection and punishment alone?

We split the question into two parts because we *do not* take norm internalization as an exogenous given. Instead, we include it as a possible agent strategy, with the possibility wide open that the strategy might go extinct (this is true of any strategy in the strategy space). Thus, to have an effect that we care about, norm internalization must (1) not go extinct and (2) influence cooperation levels

positively, which is by no means a given, since norm internalizers may internalize a defection norm (for more reasons see (Henrich and Boyd 2001; Lehmann et al. 2008a)). This means that we are concerned both with the selective pressures in favor of norm internalization (its causes, Q1), and its consequences (Q2).

The reason norm internalization could have an impact above the already powerful effects of punishment (Boyd et al. 2010; Boyd et al. 2003; Boyd and Richerson 1992; Henrich and Boyd 2001) is that punishment, in our model, is imperfect: many defections go undetected and, furthermore, agents know how likely they are to be caught. In these cases, can norm internalization do what punishment on its own cannot: keep agents cooperating even when no one is looking?

Most previous work on the issue, in contrast to ours, (1) does not include group selection or punishment (and certainly not the imperfect punishment we use) (Lehmann and Feldman 2008), (2) holds the norm internalization trait fixed, therefore not answering question (Q1) (Henrich and Boyd 2001; Lehmann et al. 2008a), (3) uses a non-conformist norm internalization mechanism, where the pressure to internalize a norm is instead specified exogenously (Gavrilets and Richerson 2017; Lozano et al. 2020), or (4) assumes some outside benefit to conformism (Gintis 2003).

We hypothesized that norm internalization would increase between-group variation, thereby amplifying group selective forces (Boyd and Richerson 2005; Henrich 2004). This would then favor cooperative groups, which we expected would contain more NIs — positive feedback in favor of NIs. However, we found that while norm internalization polarized groups, high-cooperation groups did not necessarily have more NIs. Thus, the basic feedback hypothesis was too simple.

Regarding Q1, we found that norm internalization didn't quite proliferate when rare, as only a minority of agents possessed the trait in all simulations we ran. Instead, the population of NIs averaged between a third and almost zero. Nevertheless, with regard to Q2 we found that NIs *did* substantially increase cooperation levels, when both cooperation and norm internalization started out rare, over and above the effects of group selection and punishment. Importantly, this effect did not obtain when the NI population was the highest; rather, it was the most pronounced when NIs constituted on average about 10% of the population.

How did such a small population of NIs increase cooperation levels so substantially? They sparked higher levels of cooperation in other agents, by playing the following roles:

Polarizer A high prevalence of norm internalizers (NIs) in a group tended to lead either to extreme cooperation or extreme defection among its members, amplifying intergroup differences in cooperation levels.

Catalyst NIs were especially effective at precipitating cooperation (among all agent types) when global cooperation levels started very low.

Stabilizer The presence of NIs in groups tended to prolong bouts of high total cooperation.

We derived our results from agent-based simulations where agents (1) play a public goods game in groups and (2) evolve over generations, with cooperation always starting rare. We addressed Q1 by starting with a low NI population and observing the results — this enabled us to see whether NIs could survive without going extinct, and what population level they attained. As for Q2, we compared two conditions: versions of the simulation that included NIs in the strategy space (with nothing preventing the small starting population from going extinct) with ones that did not (starting instead with a small population of unconditional cooperators). This helped us to parse whether norm internalization was able to increase cooperation levels beyond what group selection and punishment could do alone.

We performed these tests with two very different agent-based models, choosing this approach to show that our results, which turned out to be quite similar for the two models, do not depend on specific implementation details or assumptions about, for instance, group conflict rates.

Section 2 will provide further background on norm internalization, group selection, and punishment, the three interacting 'prosocial forces' in our model. Then, Section 3 describes the models (which are fully specified in the supplementary materials), and Section 4 presents the results, whose implications are discussed in Section 5 and Section 6.

2 Background

Most of the literature on norm internalization focus on its consequences for the evolution of cooperation, rather than on its causes. Here, we are concerned with both the causes and consequences of norm internalization, with Q1 concerned with the former and Q2, the latter. In this section, we first define norm internalization and review the literature on its causes (the selective pressures favoring it) and consequences. We also examine how the two other forces in favor of cooperation, group selection and punishment, interact with norm internalization.

2.1 Defining norm internalization

We define norm internalization as the tendency to follow the majority behavior of the community, even at the expense of one's own fitness. This is closely linked to the notion of conscience, which is an internal enforcer of norm-following that is powered by emotions such as guilt and shame (Frith and Metzinger 2016; Tangney 2005). For instance, Boehm describes conscience as the "internalization of values" (Boehm 2012), and Churchland acknowledges that conscience involves "feelings that urge us in a general direction, and judgment that shapes the urge into a specific action" (Churchland 2020). However, because 'conscience' is bound up with moral emotions that are not explicitly represented in our models, we will stick to the term 'norm internalization'.

Empirically, humans internalize norms from those around them (Grusec and Goodnow 1994; Parsons 1967), and follow norms even when not being observed (Fischbacher and Föllmi-Heusi 2013). Children, furthermore, are 'promiscuous normativists', ascribing a normative character indiscriminately to observed actions, even going so far as to enforce behavior in accordance with those actions (Schmidt et al. 2016). In what follows, we aim to elucidate the evolutionary origins of these empirically observed traits.

2.2 Possible causes of norm internalization

Norm internalization as described in Section 2.1 is a type of *conformist transmission* (defined as copying prevalent behaviors in the group (Henrich 2004)), which Henrich and Boyd showed to be beneficial in a noisy environment, as it allows individuals to effectively base their decisions on a large number of samples from the environment – that is, the samples taken by everybody else, and not just their own. Because a larger set of samples better approximates the ground truth, conformist behaviors can lead to better-adapted outcomes (Henrich and Boyd 1998). Coordination games also provide selection pressure in favor of conformist transmission (McElreath et al. 2003; Young 1998). This beneficial aspect of conformism has led to the exaptation hypothesis, which states that humans developed a propensity to copy others because it is usually beneficial, but individually-harmful altruistic behaviors are also copied 'by mistake' (in evolutionary terms) (Gintis 2003; Henrich and Boyd 2001). Still, copying behaviors would result in a net benefit for fitness, given the quantity of cultural behaviors that people simply could not learn on their own (Henrich and McElreath 2003). The exaptation hypothesis has a shortcoming. Many altruistic actions are manifestly costly, in terms of time (participating in group hunts), forgone benefits (not stealing), or a risk to one's life (going to war). For this reason, the assumption that an agent would learn to copy behaviors indiscriminately as a noise-reduction mechanism seems too strong: more plausibly, conformist transmission helps the agent decide among multiple options about which some information is available. Thus, in this paper, we don't assume any external benefits to the norm internalization trait (such as the ability to learn from others in a noisy environment).

Another proposed cause of the tendency to abide by the cooperation norm, even when there would be no future consequences to defection, is *risk-management*. In a world full of agents using the tit-for-tat strategy (Axelrod 1981), it is much more costly to defect when one should have cooperated (since it leads to an endless chain of future defections) than it is to cooperate when one could have defected (since this simply involves a one-time payment of the cooperation cost) (Delton et al. 2011). This is an elegant explanation of norm-following under certain conditions, but it is not quite

sufficient for the scenario we describe, for two reasons. First, it assumes a world of tit-for-taters, when there is no particular reason for doing so, since tit-for-tat is not an evolutionarily stable strategy (no pure strategies are stable in the iterated prisoner's dilemma (Lorberbaum 1994)). Second, in n-person interactions, which is what we examine here with a public goods game, direct reciprocity in the style of tit-for-tat quickly collapses (Boyd and Richerson 1988).

2.3 Group selection: Consequences of norm internalization and conformist transmission

Both norm internalization and conformist transmission have an important consequence: they can amplify group selection. Group selection is vital in sustaining cooperation (Wilson et al. 2008), because it can offset the individual's loss from cooperation with benefits accrued by the group. The strength of group selection depends on the variance among groups with respect to the trait in question (Henrich 2004), and as a result, it is particularly effective when groups are small, so that they deviate more from the average by the law of large numbers; when migration rates are low, so that differences among groups are maintained (Maynard Smith 1964); and, importantly, when conformist transmission is strong (imagine two groups respectively with 40/60 and 60/40 cooperation/defection ratios – if conformist transmission occurs, these groups will respectively move toward full defection and full cooperation as agents copy the majority behavior, increasing inter-group variation (Henrich and Boyd 1998)). Thus, conformist transmission, with this polarizing tendency, could boost the evolution of cooperation (Boyd and Richerson 2005; Fehr et al. 2002; Henrich 2004), though with some caveats (Lehmann et al. 2008a).

2.4 Co-evolutionary hypothesis as a cause of norm internalization

This polarizing consequence of norm internalization leads us to the hypothesis that norm internalization becomes entrenched by amplifying group selection and co-evolving with cooperation. Several studies have explored this hypothesis, with mixed results:
Gavrilets and Richerson found that NIs evolved when there is strong pressure to punish free-riders in

part because, by cooperating, they allow the group to save on the cost of punishment (Gavrilets and Richerson 2017). However, their version of norm internalization was not dependent on the frequency of the trait — instead, the social pressure was exogenously determined by a model parameter. Our paper differs in two important respects: first, their punishment mechanism (which was central to their result) was perfect, in that all defectors could be punished. In our study, we are interested in the possibility that norm internalizers might fill the gap of imperfect punishment, where not all agents are caught, and thus we focus on the internalization of the *cooperation* norm — which could potentially keep agents cooperating even when no potential punishment is observing them. Further, while Gavrilets et al (and other more recent papers, such as (Lozano et al. 2020)) study a version of norm internalization whose strength is specified by an exogenous parameter, our experiment examines the case where the social pressure to follow the norm depends on the percentage of group members that follow it — leading to very different dynamics.

Lehmann and Feldman found that generally, conformist transmission of helping behaviors (which can be interpreted as norm internalization) does not promote culturally transmitted cooperation (Lehmann and Feldman 2008). One reason for that could have been the model's assumption that groups completely reshuffle every generation, which severely hampers the development of inter-group differences. In contrast, in this paper, we relax this assumption, and also include punishment by indirect reciprocity.

2.5 Punishment by indirect reciprocity

With group selection, norm internalization has an amplification effect, and with punishment, it has a complementary effect. Norm internalization is cheap when cooperation is rare, because (costly) cooperative norms are only internalized when a majority of agents cooperate. In comparison, punishment is expensive when cooperation is rare: having many defectors means that the population must expend more resources on meting out punishment (Boyd et al. 2003) (see (Boyd et al. 2010) for a possible solution). The inverse is true, too: norm internalization is expensive when cooperation is common (as it leads to the internalization of a costly cooperation norm), while punishment is cheap

(few defectors to punish). Furthermore, norm internalization can get agents to cooperate where the threat of punishment cannot: that is, when an action is unlikely to be observed, and therefore unlikely to be punished. Taken together, these contrasting characteristics suggest that norm internalization and punishment would effectively complement each other.

In our model, the punishment mechanism is an imperfect variant of indirect reciprocity, namely, the withholding of benefits from those caught defecting — imperfect because not all defections are observed (cf. (Hirshleifer and Rasmusen 1989)). This form of punishment is not costly for the punisher, and it creates an incentive for egotistic agents to cooperate at least some of the time. This practice is evolutionarily stable under certain conditions (Odouard and Price 2023; Ohtsuki and Iwasa 2006; Panchanathan and Boyd 2003), is observed in many societies (Bhui et al. 2019; Henrich and Henrich 2014), and is reproduced in experiments (Wedekind and Milinski 2000). For all these reasons, we *fix* the punishment mechanism, focusing not on the co-evolution of punishment and norm internalization, but rather on whether norm internalization can do its job, given the existence of an imperfect punishment mechanism.

2.6 Our niche

In brief, then, in contrast to Lehmann et al. and Gavrilets and Richerson, our study focused on the interaction between frequency-based NIs, group selection, and 'imperfect' punishment. Further, no external benefits to agents are assumed (contra (Gintis 2003)), and no particular strategy is assumed (contra (Delton et al. 2011)). Could this combination of factors lead both norm internalization and cooperation to proliferate when rare?

3 The models

To address our questions, we designed agent-based models with free-floating populations of various agent strategies, in which both norm internalization and cooperation started rare. No version of the model had a fixed distribution of strategies — their relative frequencies were always free to evolve. Our objective was to measure (Q1) the equilibrium population of norm internalizers (NIs) — did

they go extinct? sweep to fixation? and (Q2) the differences in cooperation levels when NIs were and were not present in the strategy space (allowing us to measure the effect of NIs over and above group selection and punishment).

We tested two models[†] because agent-based simulations can yield very different results with only minor differences in implementation (see (Galan and Izquierdo 2005; Huberman and Glance 1993; Nowak and May 1992)). The first, 'abstract model' bears many similarities to previous work (Boyd et al. 2003; Nowak 2006): co-location of groups, fixed group sizes, fixed numbers of groups, etc. This makes it easier to both compare it with other models. The second, 'naturalistic model' is more complex, locating groups in space, inspired by (Grimm et al. 2005). We give an overview of the model in this section; the appendix (in the electronic supplementary materials) contains a full specification.

3.1 Basics

Both models follow the same series of steps:

- *Decision* — Agents decide, based on their strategy, whether to cooperate or defect.
- *Distribution* — Public benefits are distributed to agents that were not caught defecting (punishment occurs in this step).
- *Inter-group dynamics* — Groups compete, either indirectly for resources or directly by conflict. Migration occurs.
- *Selection* — Individuals survive and reproduce according to fitness.

3.1.1 Decision

Agents, divided into groups, decide whether to cooperate or defect in a public goods game (Boyd et al. 2003). To cooperate means to pay a cost, c, to produce a benefit, $b > c$, that will be shared among the group (Talhelm et al. 2014). Agents also have a probability of being observed, which is sampled uniformly at random, for each agent, every round. This probability can influence agent decisions.

[†]Code and data at `https://github.com/victorvikram/norm-internalization-and-coop`

Agents decide whether to cooperate using a strategy parameterized by two variables: (1) the propensity to cooperate, π, and (2) the learning style, which defines how π changes. The different values of π produce a continuous space that includes unconditional defection, expected value (EV) maximization, and unconditional cooperation. Norm internalization is encoded in the learning style. More specifically, with p_{obs} being the probability of being observed and $\bar{b}$ the average group benefit distribution (see Section 3.1.2) in the previous round, an agent cooperates when

$$p_{\text{obs}}\bar{b} \geq (1 - \pi)c, \tag{1}$$

unless they err, with probability ϵ. Thus, $\pi = 0$ corresponds to an EV-maximizer, as the resulting equation compares the expected cost of defecting ($p_{\text{obs}}\bar{b}$) to the expected cost of cooperating (c) and picks the lower-cost option. By contrast, $\pi = 1$ corresponds to unconditional cooperation (since $p_{\text{obs}}\bar{b} \geq 0$ no matter what), and $\pi \ll 0$ corresponds to unconditional defection. The learning style defines how π changes:

- *Norm internalization* - Gradually approach a π value of 1 (cooperation regardless of who is watching) as long as a majority of other agents cooperate; otherwise, approach 0 (EV-maximization).
- *Selfish* - Move in the direction that yields the best individual payoff.
- *Static* - Don't learn.

We included the selfish learners to strengthen our results, as individual learning likely hampers the evolution of cooperation (Lehmann et al. 2008b). Further, it ensures NIs are not be the only agents with a dynamic π-value, providing them with adequate competition from a selfish agent that also had a dynamic π-value. Of course, many selfish learning rules are possible, we chose this one for simplicity — our focus is to compare the presence and absence of NIs in the strategy space, not the effects of different types of selfish strategies.

3.1.2 Distribution

After agents make their choices, they are observed with probability p_{obs}. The total contributions of cooperators are distributed to group members who were not observed defecting, each receiving a share $\overline{b}$. Thus, agents for whom there was either (1) a high p_{obs} or (2) a high $\overline{b}$ have a stronger incentive to cooperate. The distribution of the benefits results in the payoffs shown in Table 1 (see also Fig. 1.a2,n3):

3.1.3 Inter-group dynamics and selection

Inter-group conflict is implemented quite differently in the two models, as described in detail in the following Sections 3.2 and 3.3. As for migration, we test a variety of migration rates, but in both models focus most of our analysis on parameters that yield approximately a 50% migration rate – that is, agents have a 50% probability of dying in a group that they were not born in, motivated by the high migration probabilities observed in hunter-gatherer societies (Hill et al. 2011).

3.1.4 Parameters

We ran both models on the lower end of the range of values in which cooperation did not go extinct, which was $b/c \in [2.75, 4.25]$, with increments of 0.25 (benefits lower than this range never resulted in cooperative worlds). We tested two conditions:

- *Norm internalization* - 2% of agents started with the norm internalization trait.
- *No norm internalization* - No NIs were present in the strategy space, so we replaced the 2% starting population of NIs with unconditional cooperators.

For both conditions, the remainder of agents were split between selfish and static learners, with each with a (unique) propensity π uniformly sampled between -1 and 1. We did this to make as few assumptions about the initial population as possible. That said, agents with π-values far from zero quickly died off, leaving behind, at the start of the run, agents who were effectively EV-maximizers.

As for why we included unconditional cooperators in the non-NI model, we needed a cooperative-type of agent for forces like group selection and punishment to favor, to fill the role of the NIs in the NI-condition. The obvious choice to fill this role is the unconditional cooperator. The key parameters appear in Table 2. A full list appears in the appendix.

3.2 Abstract model

In the abstract model, the decision and distribution steps are exactly as described in Section 3.1. Regarding intergroup dynamics, groups pair up (Fig. 1.a3), and engage in conflict with probability p_{con}. Groups with higher average fitness are more likely to win (Fig. 1.a4). Following Bowles, we set p_{con} to approximately one conflict every four agent lifetimes (Bowles 2001).
Then, in the migration step, individual agents pair up and switch groups with probability p_{mig} (Fig. 1.a5).
As for the (individual) selection phase, a fixed number of agents survive and reproduce every round, chosen probabilistically according to their fitness (Fig. 1.a6).

3.3 Naturalistic model

In contrast to the abstract model, the naturalistic model locates groups on a spatial grid. The spatial aspect causes groups to compete by encroaching on each other's foraging grounds, obviating the need for a direct conflict component (Fig. 1).

3.3.1 Decision

In addition to the decision on whether to cooperate (Section 3.1.1), agents also decide where to forage (Fig. 1.n1), knowing that the more foragers there are on a square, the lower their payoff (due to competition; Section 3.3.2). An agent calculates the expected payoff of foraging on their home square compared to that of foraging on a randomly chosen adjacent square and chooses where to go with probabilities proportional to these payoffs (there is an additional "cost of foraging on distant lands," c_{dist} so even if one's home square is more crowded, it may still pay to stay put).

3.3.2 Distribution

The only addition to what is described above is that an agent's payoff is scaled by the number of agents foraging on the grid square. That is, both the private and public benefit generated by an agent are divided by n_{here}, the number of foragers on that agent's square.

3.3.3 Intergroup dynamics

In this model, groups compete for resources without direct conflict. Groups grow when their members have high fitness (Section 3.3.4), leading to crowding, so there is a mechanism by which a new group can bud off an old group (Fig. 1.n6). Recall that agents need not necessarily forage on their group's home square (Section 3.3.1). Generally, when they forage on another square, they remain tied to their group. However, if at least n agents from a group are foraging on a square other than their home, and they make up the plurality of agents on that square, then they start a new group on that square if none is already there. In this manner, populous groups can spread across the landscape. Because the benefit derived from foraging is inversely proportional to the number of agents present, these new groups begin to compete for the resources of pre-existing groups. Groups that make the most of available resources (the cooperative ones) are therefore expected to out-compete the others. Finally, if an agent is foraging away from their group's home square, and there happens to be another group on their current square, they may migrate to that group with probability p_{mig} (Fig. 1.n2).

3.3.4 Selection

In this model, there are costs of surviving and costs of reproducing. An agent's ability to pay the cost of surviving (resp. reproducing) determines whether they survive (resp. reproduce), as shown in Fig. 1.n5. For a group to grow, it must have a significant proportion of agents with high enough fitness to pay both costs. Thus, high group-average fitness leads to growth, in contrast to the abstract model, where group size is constant and high group-average fitness leads to an advantage in direct conflict.

4 Results

Despite all the differences in implementation between the abstract and naturalistic models, both models exhibited remarkably similar results. In what follows, we will flag qualitatively different results with a *difference* annotation. Further, whenever there is a figure that is specific to one of the models, the appendix contains the counterpart for the other model. In both models,

1. norm internalizers (NIs) made up a small minority of the population. In fact, for the parameter settings under which they made the biggest difference in cooperation levels, they constituted an average of 10% of the population.
2. NIs had bouts of high cooperation that were significantly *shorter* than those of the other agents (though they did go higher).
3. The presence of NIs did not lead to higher levels of cooperation in groups in which they were more common

Yet, we found that for mid-range benefit-to-cost ratios, the norm internalizer condition had

1. *higher* levels of cooperation
2. *longer* peaks of above-average cooperation.
3. fewer long bouts of below-average cooperation (though this held only in the naturalistic model).

To shed light on this counterintuitive result, we show that norm internalizers

1. *polarize* groups either to extreme cooperation or extreme defection, perhaps enhancing group selective forces that favor cooperative groups;
2. tend to be the ones to *catalyze* bouts of above-average cooperation in all agents – not just other NIs – when cooperation is especially low (this result was stronger in the naturalistic model);
3. help to *stabilize* high levels of cooperation, keeping peaks high for longer.

4.1 Average cooperation in each condition

The presence of norm internalizers (NIs) in the strategy space (the NI-condition) either increased mean levels of cooperation or made no difference. In Fig. 2, we show the range of benefit-to-cost (b:c) ratios for which the presence of NIs did make a difference: the mid-range. That is, there is a low range of b:c ratios for which surplus cooperation is essentially zero (surplus cooperation is the amount of cooperation above the 5% error rate), and a high range for which relatively high cooperation levels emerge, regardless of whether NIs are present. But for ratios in between those extremes — from at least `3 - 3.5` in the naturalistic model and `3.25 - 3.5` in the abstract model — NIs made a significant difference. In the majority of these mid-range conditions, their presence at least doubled the surplus cooperation.

Despite their tendency to boost cooperation, NIs did not necessarily cooperate more than other agent types. Even if they did, most of the boost in average cooperation was due to NIs catalyzing high cooperation among *other* agent types: in Fig. 2, the majority of the difference in total cooperation in the NI condition (medium dots) and the non-NI condition (dark dots) is accounted for by the rise in cooperation of non-NIs.

In fact, in the naturalistic model, the NI-condition boosted cooperation even when NIs tended to cooperate *less* than other agent types (with $b/c = 3$), and the NI condition made the most difference when NIs cooperated about equally ($b/c = 3.5$). In the abstract model, however, NIs always cooperated more on average than their counterparts (one notable *difference* between the models). That said, both models shared the characteristic that NIs had a higher variance in their cooperation levels, which can be seen in Fig. 3a.

The drop-off in the NI's ability to catalyze cooperation for larger b:c ratios likely has to do with with the drop in their population as b:c rises. This is because, as b:c rises, the extent to which NIs cooperate disproportionately also grows (see the light dots in Fig. 2). This causes their population to take a hit (see Fig. 3b). With that said, at no b:c did NIs make up a majority of the population (averaged across rounds). In fact, their population was only around 10% for the ratios at which they

were *most* effective at bringing about higher levels of cooperation (3.25 in the naturalistic model and 3.5 in the abstract model).

4.2 Shape of cooperation in each condition

The average only tells part of the story, however, especially because in our simulations, levels of cooperation tended to rise and fall cyclically. Two factors, therefore, could have increased the average: either the NI condition had relatively longer peaks (compared to its troughs) or it had higher average peak (and/or trough) cooperation levels.

As can be seen in Fig. 4 (on the right), the NI-condition exhibited both properties. That is, the peaks were longer and they had higher levels of cooperation. Further, a larger portion of the round was spent in the peak, but this difference was not significant. We define a peak as an interval in which cooperation is above average for at least ten rounds, and a trough as the complement of that.

This result stands in contrast to the cooperation peaks of the NIs themselves (at left). NI cooperation peaks, while occupying about the same proportion of the round, were much shorter. And while NIs cooperated more on their peaks, they cooperated less in their troughs, which meant that NIs cooperated barely (and certainly not significantly) more than non-NIs overall (*difference*: in the abstract model, NIs did cooperate more).

4.3 The roles of norm internalizers

We are left, then, with a puzzle. NIs themselves do not exhibit longer peaks of cooperation, nor do they necessarily cooperate more on average than other types of agents. The presence of NIs in the population must, therefore, contribute to longer peaks and higher cooperation levels indirectly. Indeed, as the rest of this section shows, NIs play three vital roles in facilitating cooperation.

4.3.1 Polarize

The central role of the norm internalizer is that of *polarization*. High levels of NIs in a group caused cooperation to cluster at the extremes: either very low or very high. This makes sense: when

cooperation is high, NIs internalize the tendency to cooperate, further increasing cooperation levels. The opposite is true at low levels of cooperation. We address in the discussion the intimate connection that polarization has with catalysis and stabilization (Sections 5.3 and 5.4).

Polarization provides the first hint as to why NIs themselves do not have higher average cooperation levels: norm internalizers exhibit more extreme, not necessarily more cooperative behavior. Indeed, Fig. 5 shows that higher populations of norm internalizers in a group do not increase the mean level of cooperation. This effect — that high norm internalizer populations do not imply higher cooperation — is the converse of the phenomenon observed earlier — that higher cooperation does not imply higher norm internalizer populations (Fig. 3b).

4.3.2 Catalyze

The second role of the norm internalizer is its ability to catalyze cooperation, allowing it to bootstrap itself from extremely low starting points. This role was much more pronounced in the naturalistic than the abstract model (an important *difference* between the models). We will propose reasons for why this might be so in Section 5.3; for now, we focus on the evidence from the naturalistic model.

As can be seen in Fig. 6, there are more long troughs in the non-NI condition. For the NI condition, if a point is in a trough, there is a 3% probability that it is a trough of more than 3,000 rounds, while for the non-NI condition, that number is 33%. We sought to elucidate this result by looking at which agent type *initiated* cooperation spikes – that is, the agent type that exceeded its mean cooperation level first. This result is plotted in Fig. 6(d, e), which shows that for both longer and deeper troughs, NIs played a disproportionate role in initiating subsequent spikes. This helped prevent the NI-condition runs from getting stuck in very long troughs.

A useful way to examine this phenomenon is with a state-transition diagram, shown in Fig. 7. While the transitions between these particular states are not technically Markovian, they nonetheless show the likely successor for each state. There are two things to observe here. First, the most stable states (the ones with the greatest proportion of self loops) are those with very low cooperation (top left), and those with above average cooperation of both NIs *and* non-NIs (bottom right). Effective

'catalysis' of cooperation, then, corresponds to moving from the top left to the bottom right. Notice that, in the naturalistic model, one of the most likely paths that achieves this is the one from very low cooperation (upper left) to high-NI cooperation (top right) to high cooperation for all agents (bottom right). All other two- or three- step paths have a much lower probability. The advantage of this path is much less pronounced in the abstract model (left), which is consistent with our observation that NIs don't play as much of a catalyst role in that setting.

4.3.3 Stabilize

While the catalyst role was much stronger in the naturalistic model, the third, stabilizing type of effect that NIs had on high levels of cooperation (reflected in the longer peak lengths in the NI condition, along with the higher average levels of cooperation) was strongly present in both. This may seem counter-intuitive, because the peaks of NIs themselves are shorter than those of other agents. The key point, however, is that when the NI population is higher than average, cooperation peaks among *all* agent types, not just NIs, tends to be maintained (right side of Fig. 4).

We examine this effect by zeroing in on individual groups and their cooperation patterns. In particular, we first examine the regression coefficient between past and future levels of cooperation in high-NI vs. low-NI groups. We found that the coefficient is much closer to 1 in high-NI groups, which implies a much better sticking power of cooperation (see Fig. 8). We also examined state transitions between low (<50%) and high (>50%) cooperation, finding that a high level of NIs was much more effective at maintaining high cooperation levels: 96% of the time, high cooperation groups stayed high in high-NI groups, but only 76% of the time in low-NI groups (see Fig. 9a).

While high populations of NIs are effective at maintaining high levels of cooperation, it is also true that high levels of cooperation lead to a decline in their population (see Fig. 9b). A Granger causality test of the effect of cooperation levels on norm internalizer population supported this observation: with very high confidence ($p < 0.0001$), high cooperation levels Granger-caused lower subsequent NI populations. This echoes the tendency for NI populations to fall at higher b:c ratios (Fig. 3b), when cooperation levels tend to be higher. Furthermore, while NIs tended to help high cooperation

groups stay high, they also had a small but significant role in keeping low- (under 50%) cooperation groups low (see Fig. 9a). We will discuss the apparent tension between this fact and the catalyst role of norm internalization in Section 5.3.

5 Discussion

The previous section highlighted the evidence for the roles that norm internalizers (NIs) played in raising average levels of cooperation. Here, we offer mechanistic explanations (or hypotheses), and ease some of the apparent tensions between the three roles.

5.1 Overall effect

First, we examine the overall effect of NIs on cooperation levels: when they were present *in the strategy space* — not necessarily when their population was larger — they tended to boost cooperation levels (Fig. 2). This is consistent with the result that, when selection on a trait depended on Darwinian fitness (as the helping trait does in our model), unbiased imitation can help increase cooperation levels (Lehmann et al. 2008a). The difference is, in our model, norm internalization is not an exogenous given but an evolving trait that can go extinct. In this latter respect our work is similar to (Lehmann and Feldman 2008), though we obtained different results likely because their model included neither persistent groups (they reshuffled every round) nor punishment. For NIs to have any effect in our model, both of these mechanisms were required. This echoes Gavrilets and Richerson's finding about *non-conformist* NIs (Gavrilets and Richerson 2017) they, too, had a much higher effect on behavior in the presence of punishment.

Furthermore, most of that boost came not from NIs own higher cooperation, but from increased cooperation levels of other agents: even if NIs *did* cooperate more than other agents (which was not always the case), it only made a small difference in the overall mean since their population was so small. Importantly, the three roles of NIs discussed earlier do not require higher long-term average levels of cooperation among NIs: instead, they effectively carry other agents along. As catalysts and stabilizers, NIs raised or maintained cooperation levels of other agents. And as polarizers, they

adopted (a more extreme version of) whatever agents in their group were already doing.

How do NIs carry other agents along? In our model, there is a basic mechanism accomplishing this. Those caught defecting are punished by an indirect reciprocity mechanism: the group benefit of cooperation is withheld from them. Notice that the group benefit increases as the number of cooperators in the group increases, so it becomes more and more costly to defect as cooperation becomes more common. This increasing cost of defection causes *other* agents to cooperate more often as cooperation becomes more common, even if they are not NIs. However, while the payoff difference between cooperation and defection decreases when cooperation becomes common, the individual payoff of cooperating rarely, if ever, exceeds the cost of defection. This contrasts with the structure of a stag hunt game, in which it pays to simply adopt the majority strategy.

5.1.1 Population-cooperation tradeoff

An important factor in the ability of the NI-condition to catalyze cooperation is the trade-off between population and cooperation level. The NI population has to be high enough to make a difference, but that is not enough: it is of no help for them to constitute a third of the population if their cooperation level is kept down as a result of internalizing the norm of defection (this is what happened, for instance, at a benefit-to-cost ratio (b:c) of `3.0` in the abstract model). But — and here is the catch — the more NIs cooperate, the more their population falls (Fig. 9). Thus, to raise cooperation levels, NIs must cooperate, but so much that their population falls. The presence of this tradeoff explains why NIs are most effective with mid-range b:c ratios: if b:c is too high, NIs cooperate too much, leading their population to diminish. If it is too low, NIs never internalize the norm of cooperating, so their numbers can grow to make up about a third of the population without helping to spark cooperation (see Fig. 3). These fluctuating population dynamics are in stark contrast to the results of Gavrilets and Richerson, who observe a more or less stable level of norm internalizers once the trait has caught on (Gavrilets and Richerson 2017) This is no surprise, as their version of norm internalization is not a conformist one, and should therefore be less sensitive to the population distribution of other cooperators.

The mid-range is especially important as the region where cooperation is possible, but the hardest to maintain. One may of course ask, is the added cognitive machinery required for the norm internalization strategy worth the cost if all it does is lower the threshold for appreciable amounts of cooperation to a lower b:c ratio? The answer is quite possibly yes, especially because the machinery required to imitate the behaviors of others may have already existed for copying beneficial behaviors, and could have been co-opted at minimal cost (see e.g. (Gintis 2003; Henrich and Boyd 1998)).

While the naturalistic and abstract model yielded largely similar results, there were some differences, including two major qualitative differences. First, in the abstract model, norm internalizers always cooperated more than others, while in the naturalistic model, it depended very much on the b:c ratio. Second, the NI-condition in the abstract model actually saw increasing or flat cooperation levels as migration rates increased, compared with declining ones in the spatial model. While we cannot say definitively what caused these differences, it makes sense that migration had different effects in the two models, as the spatial structure of the naturalistic model meant that only agents from neighboring groups mixed, while in the abstract model, the mixing was uniform.

We next discuss the three roles played by norm internalizers.

5.2 Polarize

We mentioned in Section 2.3 that between-group differences are necessary for group selection to be strong. Norm internalization, by polarizing groups, increases between-group differences, which in turn can enhance selection in favor of cooperation. Our result is thus in line with previous results on conformist transmission enhancing inter-group differences (Henrich 2004).

There may seem to be a tension between two findings: high-NI groups do not have higher average cooperation (Fig. 5), but nevertheless, the NI condition has higher cooperation than the non-NI condition. However, even though NIs may have an ambiguous effect on intra-group cooperation, the differences they cause *between groups* can enhance group selection in favor of cooperation. We will discuss in the next section (5.3) how the inter-group dynamics resulting from polarization could lead to catalysis, and then in the following one, (5.4), how the intra-group effects of polarization can lead

to stabilization. Hence, polarization is the central role whose different manifestations help give rise to the others.

5.3 Catalyze

In their role as catalyst, NIs helped to spark spikes in cooperation when cooperation levels were especially low. In this section, we propose a mechanism by which this might have occurred, resolve apparent tensions between this function and the other two, and provide hypotheses for why this role was less salient in the abstract model.

As can be seen in Fig. 9, NIs face no selective disadvantage at low levels of cooperation. This means that when cooperation is very low, their population can grow quite large, by drift. Of course, we know that a large NI population is not sufficient to bring about cooperation, as they can internalize a defection norm. So why might this population start cooperating? The answer is that even in global states of low cooperation, there will be variation between groups: some will cooperate more, others less. And crucially — due to the polarizing tendency (here is the promised link between polarization and catalysis), and the higher variability of NI cooperation (Fig. 3) — the groups with unusually high or unusually low cooperation will probably be the ones with more NIs. Especially in environments where cooperation is globally very low, the higher variance caused by NIs is needed to produce higher-cooperation groups. These groups are favored by group selection, and their spread is what brings about a global spike in cooperation. Notice that catalysis is thus a result of the *inter-group* dynamics resulting from polarization.

Why, then, was catalysis less pronounced in the abstract model? The catalysis mechanism that we propose requires *simultaneous* inter-group variation: in the same round, there must be some higher and some lower cooperation groups. This may be much easier to achieve when there is spatial structure: the high migration rates means that groups mix quite quickly, but the spatial structure in the naturalistic model allows for geographically structured inter-group variation that is not possible in the abstract model.

This proposed mechanism helps to resolve apparent tensions between the catalyst role and the other

two. One might wonder, in particular, how NIs can be catalysts of cooperation when Fig. 9a, demonstrating stabilization, shows that it is harder for high-NI groups to go from low cooperation to high cooperation (due to the polarizing tendency of norm internalization bringing groups under 50% cooperation closer to zero). The key insight is that the figure shows the *intra-group* effect of NIs: our proposed mechanism recognizes that intra-group, NIs have an ambiguous effect on cooperation, and it is only when we consider the *inter-group* variation, along with group selection favoring cooperative groups, that the catalyst effect can work.

5.4 Stabilize

The stabilizing mechanism is more intuitive, and has two parts. First, when NIs are present in larger quantities, and cooperation is high, they will be internalizing the cooperation norm — this is simply the positive side of the polarization phenomenon. This will further increase the levels of cooperation, causing the peaks of high cooperation to stick longer. Second, as described earlier, the increases in NI cooperation can raise defection costs, causing other agents to cooperate more. Note that the mechanism proposed here is the result of the *intra-group* consequences of polarization. This stabilization effect is in line with Henrich and Muthukrishna's idea that group selection is greatly aided by intra-group dynamics that help to maintain a particular state (in this case, high cooperation) (Henrich and Muthukrishna 2021).

Why, given the positive-feedback dynamics described so far, doesn't cooperation persist indefinitely? The answer is that the positive feedback loop is embedded in a larger negative feedback loop. It is true that high cooperation leads to more internalization of cooperation and a higher cost of defection, which in turn lead to more cooperation; but high cooperation levels also tend to reduce the population of NIs, as seen in Fig. 9 (the same effect leads to a reduction in NI population with rising b:c, as described in Section 5.1.1). This is because when cooperation levels are very high, NIs become unconditional cooperators, and cooperate much more than other agent types, leading to a reduction in their population. This results in a subsequent reduction in cooperation levels, but if cooperation levels get too low, NI populations can rise once more, setting off the positive feedback of cooperation

again.

The cyclical dynamics can also help explain how NIs prolong overall peak durations while having shorter peaks themselves. When NI cooperation levels rise, they bring the cooperation of other agents up, too. However, in groups where cooperation gets too high, norm internalizer populations tend to die off; globally, this has the effect of killing the highly-cooperative NIs and preserving the non-cooperative NIs, which reduces the NI cooperation average, ending the NI cooperation peak. However, cooperation levels of other agent types, which had grown in response to an increasing cost of defection, remain high because their own high cooperation levels maintain that high cost of defection — so the cooperation peak of other agents persists. When it eventually does fall enough (though remaining relatively high), it opens the door for norm internalizer populations to rise again, and possibly start to bring cooperation levels back up. In this way, NI cooperation can spike when it is needed to preserve a peak of non-NI cooperation, but it doesn't necessarily remain high for the entire non-NI peak, thereby increasing peak lengths even though its own peak lengths are comparatively short.

5.5 Cultural Evolution

So far we discussed the immediate effects of norm internalization on cooperation dynamics. But there is also a broader consequence to its effects, namely, that it forms part of the basis for cultural evolution. Under genetic evolution, transmission occurs from parent to offspring, with variation introduced by mutation and recombination. Cultural evolution, in comparison, requires that individuals in a society imitate others, for example, through conformist transmission.

Existing explanations for conformist transmission (Gintis 2003) do not explain why individuals would copy obviously self-sacrificial behaviors (see Section 2.2), short of "myopia" or inability to tell the difference. The present research provides an alternative explanation for conformist transmission in these hard-to-explain cases. Here, group selection plays an important role, amplified by the ability of norm internalization to increase differences between groups. Our research, therefore, helps elucidate why humans didn't evolve into the fabled *Homo economicus* (Henrich et al. 2001), and

provides grounding for a broader theory of cultural transmission of altruistic and moral behaviors.

6 Conclusion

In this paper, we looked at how the presence of norm internalization in the strategy space was able to increase cooperation levels, beyond what group selection and punishment by indirect reciprocity were able to do alone. This was so even though norm internalizers (NIs) tended to make up a small portion of the population, and they didn't necessarily cooperate more than other agent types.

We were motivated by the question of whether NIs might be able to fill in the gap by a certain proportion of actions (about half) being unobserved, and therefore, unpunished. The answer was yes, but their effect was mostly indirect: a minority of NIs increased the level of cooperation of all strategies. This accords with our intuitions about the role of *conscience* — conscience is essentially norm internalization paired with emotions like guilt and shame to enforce norm-following 'from the inside'. Our intuition suggests that a conscience-like mechanism could help fill the gap when punishment is imperfect, and indeed, our results bear that out.

We identified three related roles that NIs played: they *polarized* groups, strengthening group selective forces in favor of cooperation, they *catalyzed* cooperation when global levels were especially low, and they *stabilized* bouts of high cooperation, keeping them going for longer.

We showed these roles in action, and, in Section 5, we proposed some more mechanistic explanations for each. Polarization was the hub role, which led to catalysis through *inter-group* dynamics and stabilization through *intra-group* dynamics.

For our inter-group dynamics, we used a wide range of migration parameters, and where required, empirically-backed conflict parameters. We did not address the question of how punishment by indirect reciprocity might remain stable; for this, see (Odouard and Price 2023; Ohtsuki and Iwasa 2006).

Importantly, we tested the effect of NIs over and above the effects of group selection and punishment, by looking at two conditions: one in which NIs were present in the strategy space, and one in which they were not. Given that both these forces are at play, a strategy space that includes

norm internalization yields higher cooperation than a strategy space that only has unconditional cooperators, and it can do so when both cooperation and norm internalization start out rare. As we have shown, they remain relatively uncommon, yet facilitate a dynamic that leads other agents to cooperate more and for longer.

Our work may be extended in several promising directions:

- Morality consists of much more than just cooperation in public goods games. Coordination games, hawk-dove games, and stag hunts all play into morality (Curry 2016; Curry et al. 2019). Further exploration might aim to understand the effects that norm internalization would have in each of these diverse contexts.

- While we used a continuous strategy space, it was still restricted to strategies that employed a particular function of certain input variables. Allowing for strategy optimization, for instance by making the agents capable of reinforcement learning, is a particularly interesting direction for future work.

References

Axelrod, R. (1981). The emergence of cooperation among egoists. *American Political Science Review*, *75*, 306–318. https://doi.org/10.2307/1961366

Axelrod, R., & Hamilton, W. D. (1981). The evolution of cooperation. *Science*, *211*, 1390–1396. https://doi.org/10.1126/science.7466396

Bhui, R., Chudek, M., & Henrich, J. (2019). How exploitation launched human cooperation. *Behavioral Ecology and Sociobiology*, *73*, 78. https://doi.org/10.1007/s00265-019-2667-y

Boehm, C. *Moral Origins: The Evolution of Virtue, Altruism, and Shame*. 2012. Basic Books. New York.

Bowles, S. (2001). Individual interactions, group conflicts, and the evolution of preferences. In J. Dealey & L. Ward (Eds.), *Social Dynamics* (pp. 155–190). The MIT Press.

Bowles, S., Choi, J.-K., & Hopfensitz, A. (2003). The co-evolution of individual behaviors and social institutions. *Journal of Theoretical Biology*, *223*, 135–147. https://doi.org/10.1016/S0022-5193(03)00060-2

Boyd, R., Gintis, H., & Bowles, S. (2010). Coordinated punishment of defectors sustains cooperation and can proliferate when rare. *Science*, *328*, 617–620. https://doi.org/10.1126/science.1183665

Boyd, R., Gintis, H., Bowles, S., & Richerson, P. J. (2003). The evolution of altruistic punishment. *Proceedings of the National Academy of Sciences*, *100*, 3531–3535. https://doi.org/10.1073/pnas.0630443100

Boyd, R., & Richerson, P. J. (1988). The evolution of reciprocity in sizable groups. *Journal of Theoretical Biology*, *132*, 337–356. https://doi.org/10.1016/S0022-5193(88)80219-4

Boyd, R., & Richerson, P. J. (1992). Punishment allows the evolution of cooperation (or anything else) in sizable groups. *Ethology and Sociobiology*, *13*, 171–195. https://doi.org/10.1016/0162-3095(92)90032-Y

Boyd, R., & Richerson, P. J. *The Origin and Evolution of Cultures*. 2005. Oxford University Press. Oxford ; New York.

Brown, J. S., Sanderson, M. J., & Michod, R. E. (1982). Evolution of social behavior by reciprocation. *Journal of Theoretical Biology*, *99*, 319–339. https://doi.org/10.1016/0022-5193(82)90008-X

Churchland, P. S. *Conscience: The Origins of Moral Intuition*. 2020. W.W. Norton & Company. New York.

Curry, O. S. (2016). Morality as cooperation: A problem-centred approach. In T. K. Shackelford & R. D. Hansen (Eds.), *The Evolution of Morality* (pp. 27–51). Springer International Publishing. https://doi.org/10.1007/978-3-319-19671-8_2

Curry, O. S., Mullins, D. A., & Whitehouse, H. (2019). Is it good to cooperate? Testing the theory of morality-as-cooperation in 60 societies. *Current Anthropology*, *60*, 47–69. https://doi.org/10.1086/701478

Darwin, C. *The Descent of Man. Volume 1*. 2011. Pacific Publishing Studio. Seattle, Washington.

Delton, A. W., Krasnow, M. M., Cosmides, L., & Tooby, J. (2011). Evolution of direct reciprocity under uncertainty can explain human generosity in one-shot encounters. *Proceedings of the National Academy of Sciences*, *108*, 13335–13340. https://doi.org/10.1073/pnas.1102131108

Fehr, E., Fischbacher, U., & Gächter, S. (2002). Strong reciprocity, human cooperation, and the enforcement of social norms. *Human Nature*, *13*, 1–25. https://doi.org/10.1007/s12110-002-1012-7

Fischbacher, U., & Föllmi-Heusi, F. (2013). Lies in disguise: An experimental study on cheating. *Journal of the European Economic Association*, *11*, 525–547. https://doi.org/10.1111/jeea.12014

Frith, C. D., & Metzinger, T. (2016). What's the use of consciousness? In A. K. Engel, K. J. Friston, & D. Kragic (Eds.), *The Pragmatic Turn* (pp. 193–214). The MIT Press. https://doi.org/10.7551/mitpress/9780262034326.003.0012

Galan, J. M., & Izquierdo, L. R. (2005). Appearances can be deceiving: Lessons learned re-implementing axelrod's 'evolutionary approach to norms'. *Journal of Artificial Societies and Social Simulation*, *8*.

Gavrilets, S., & Richerson, P. J. (2017). Collective action and the evolution of social norm internalization. *Proceedings of the National Academy of Sciences*, *114*, 6068–6073. https://doi.org/10.1073/pnas.1703857114

Gintis, H. (2003). The hitchhiker's guide to altruism: Gene-culture coevolution, and the internalization of norms. *Journal of Theoretical Biology*, *220*, 407–418. https://doi.org/10.1006/jtbi.2003.3104

Grimm, V., Revilla, E., Berger, U., Jeltsch, F., Mooij, W. M., Railsback, S. F., Thulke, H.-H., Weiner, J., Wiegand, T., & DeAngelis, D. L. (2005). Pattern-oriented modeling of agent-based complex systems: Lessons from ecology. *Science*, *310*, 987–991. https://doi.org/10.1126/science.1116681

Grusec, J. E., & Goodnow, J. J. (1994). Impact of parental discipline methods on the child's internalization of values: A reconceptualization of current points of view. *Developmental Psychology*, *30*, 4–19. https://doi.org/10.1037/0012-1649.30.1.4

Hamilton, W. (1964). The genetical evolution of social behaviour. I. *Journal of Theoretical Biology*, *7*, 1–16. https://doi.org/10.1016/0022-5193(64)90038-4

Henrich, J., & Boyd, R. (1998). The evolution of conformist transmission and the emergence of between-group differences. *Evolution and Human Behavior*, *19*, 215–241. https://doi.org/10.1016/S1090-5138(98)00018-X

Henrich, J. (2004). Cultural group selection, coevolutionary processes and large-scale cooperation. *Journal of Economic Behavior & Organization*, *53*, 3–35. https://doi.org/10.1016/S0167-2681(03)00094-5

Henrich, J., & Boyd, R. (2001). Why people punish defectors. *Journal of Theoretical Biology*, *208*, 79–89. https://doi.org/10.1006/jtbi.2000.2202

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., & McElreath, R. (2001). In search of homo economicus: Behavioral experiments in 15 small-scale societies. *American Economic Review*, *91*, 73–78. https://doi.org/10.1257/aer.91.2.73

Henrich, J., & Henrich, N. (2014). Fairness without punishment: Behavioral experiments in the Yasawa Island, Fiji. In J. Ensminger & J. Henrich (Eds.), *Experimenting with Social Norms: Fairness and Punishment in Cross-Cultural Perspective*. Russell Sage Press.

Henrich, J., & McElreath, R. (2003). The evolution of cultural evolution. *Evolutionary Anthropology: Issues, News, and Reviews*, *12*, 123–135. https://doi.org/10.1002/evan.10110

Henrich, J., & Muthukrishna, M. (2021). The origins and psychology of human cooperation. *Annual Review of Psychology*, *72*, 207–240. https://doi.org/10.1146/annurev-psych-081920-042106

Hill, K. R., Walker, R. S., Božičević, M., Eder, J., Headland, T., Hewlett, B., Hurtado, A. M., Marlowe, F., Wiessner, P., & Wood, B. (2011). Co-residence patterns in hunter-gatherer societies show unique human social structure. *Science*, *331*, 1286–1289. https://doi.org/10.1126/science.1199071

Hirshleifer, D., & Rasmusen, E. (1989). Cooperation in a repeated prisoners' dilemma with ostracism. *Journal of Economic Behavior & Organization*, *12*, 87–106. https://doi.org/10.1016/0167-2681(89)90078-4

Huberman, B. A., & Glance, N. S. (1993). Evolutionary games and computer simulations. *Proceedings of the National Academy of Sciences*, *90*, 7716–7718. https://doi.org/10.1073/pnas.90.16.7716

Lehmann, L., & Feldman, M. W. (2008). The co-evolution of culturally inherited altruistic helping and cultural transmission under random group formation. *Theoretical Population Biology*, *73*, 506–516. https://doi.org/10.1016/j.tpb.2008.02.004

Lehmann, L., Feldman, M. W., & Foster, K. R. (2008a). Cultural transmission can inhibit the evolution of altruistic helping. *The American Naturalist*, *172*, 12–24. https://doi.org/10.1086/587851

Lehmann, L., Foster, K. R., Borenstein, E., & Feldman, M. W. (2008b). Social and individual learning of helping in humans and other species. *Trends in Ecology & Evolution*, *23*, 664–671. https://doi.org/10.1016/j.tree.2008.07.012

Lorberbaum, J. (1994). No strategy is evolutionarily stable in the repeated prisoner's dilemma. *Journal of Theoretical Biology*, *168*, 117–130. https://doi.org/10.1006/jtbi.1994.1092

Lozano, P., Gavrilets, S., & Sánchez, A. (2020). Cooperation, social norm internalization, and hierarchical societies. *Scientific Reports*, *10*, 15359. https://doi.org/10.1038/s41598-020-71664-w

Maynard Smith, J. (1964). Group selection and kin selection. *Nature*, *201*, 1145–1147. https://doi.org/10.1038/201145a0

McElreath, R., Boyd, R., & Richerson, P. J. (2003). Shared norms and the evolution of ethnic markers. *Current Anthropology*, *44*, 122–130. https://doi.org/10.1086/345689

Nowak, M. A. (2006). Five rules for the evolution of cooperation. *Science*, *314*, 1560–1563. https://doi.org/10.1126/science.1133755

Nowak, M. A., & May, R. M. (1992). Evolutionary games and spatial chaos. *Nature*, *359*, 826–829. https://doi.org/10.1038/359826a0

Nowak, M. A., & Sigmund, K. (2005). Evolution of indirect reciprocity. *Nature*, *437*, 1291–1298. https://doi.org/10.1038/nature04131

Odouard, V. V., & Price, M. H. (2023). Tit for tattling: Cooperation, communication, and how each could stabilize the other. *Evolution and Human Behavior*, *44*, 359–372. https://doi.org/10.1016/j.evolhumbehav.2023.06.002

Ohtsuki, H., & Iwasa, Y. (2006). The leading eight: Social norms that can maintain cooperation by indirect reciprocity. *Journal of Theoretical Biology*, *239*, 435–444. https://doi.org/10.1016/j.jtbi.2005.08.008

Panchanathan, K., & Boyd, R. (2003). A tale of two defectors: The importance of standing for evolution of indirect reciprocity. *Journal of Theoretical Biology*, *224*, 115–126. https://doi.org/10.1016/S0022-5193(03)00154-1

Parsons, T. *Sociological Theory and Modern Society*. 1967. Free Press. New York, NY.

Queller, D. C. (1992). A general model for kin selection. *Evolution*, *46*, 376–380. https://doi.org/10.1111/j.1558-5646.1992.tb02045.x

Schmidt, M. F. H., Butler, L. P., Heinz, J., & Tomasello, M. (2016). Young children see a single action and infer a social norm: Promiscuous normativity in 3-year-olds. *Psychological Science*, *27*, 1360–1370. https://doi.org/10.1177/0956797616661182

Talhelm, T., Zhang, X., Oishi, S., Shimin, C., Duan, D., Lan, X., & Kitayama, S. (2014). Large-scale psychological differences within china explained by rice versus wheat agriculture. *Science*, *344*, 603–608. https://doi.org/10.1126/science.1246850

Tangney, J. P. (2005). The self-conscious emotions: Shame, guilt, embarrassment and pride. In T. Dalgleish & M. J. Power (Eds.), *Handbook of Cognition and Emotion* (pp. 541–568). John Wiley & Sons, Ltd. https://doi.org/10.1002/0470013494.ch26

Trivers, R. L. (1971). The evolution of reciprocal altruism. *The Quarterly Review of Biology*, *46*, 35–57. https://doi.org/10.1086/406755

Wedekind, C., & Milinski, M. (2000). Cooperation through image scoring in humans. *Science*, *288*, 850–852. https://doi.org/10.1126/science.288.5467.850

Table 1 **Payoffs in the model**. The f is the baseline fitness level that all agents receive; agents pay c (the cost of cooperation) if they choose to cooperate. The group distribution, $\bar{b}$, depends on how many agents in the group decided to cooperate: it is calculated as the total public benefit divided by the number of agents who received a share

	Defect	Cooperate
Observed	f	$f - c + \bar{b}$
Unobserved	$f + \bar{b}$	$f - c + \bar{b}$

Wilson, D. S., Van Vugt, M., & O'Gorman, R. (2008). Multilevel selection theory and major evolutionary transitions: Implications for psychological science. *Current Directions in Psychological Science*, *17*, 6–9. https://doi.org/10.1111/j.1467-8721.2008.00538.x

Young, H. P. (1998). Social norms and economic welfare. *European Economic Review*, *42*, 821–830. https://doi.org/10.1016/S0014-2921(97)00138-4

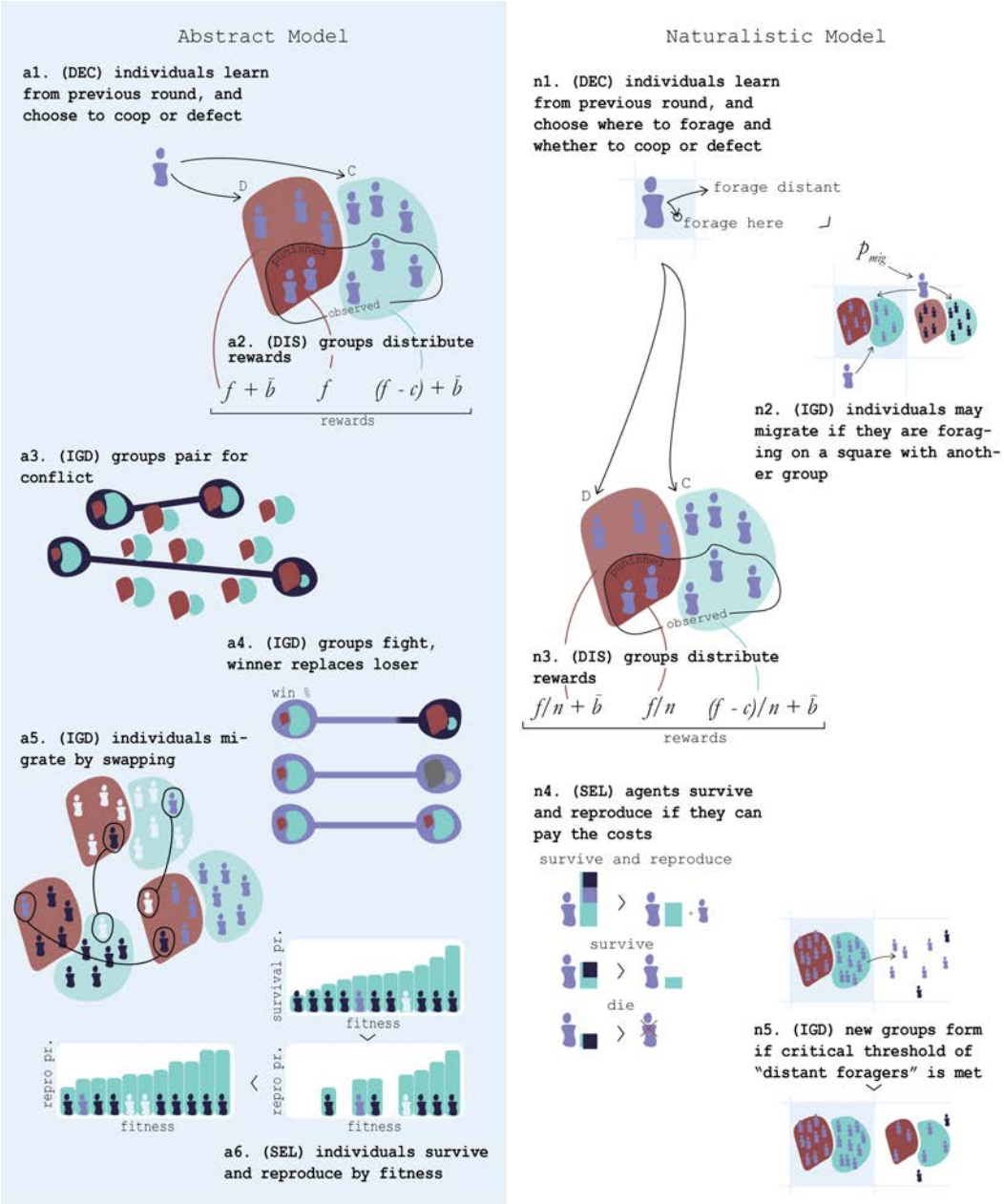

Fig. 1 **The steps taken in each iteration of the model.** Described fully in the text. Listed variables next to an in-arrow are inputs to a particular step of the model. Arrows between steps indicate what information is passed from one stage to another. In the abstract model, decision (DEC), distribution (DIS), inter-group dynamics (IGD), and selection (SEL) occur in sequence. By contrast, in the naturalistic model these steps are interspersed with each other

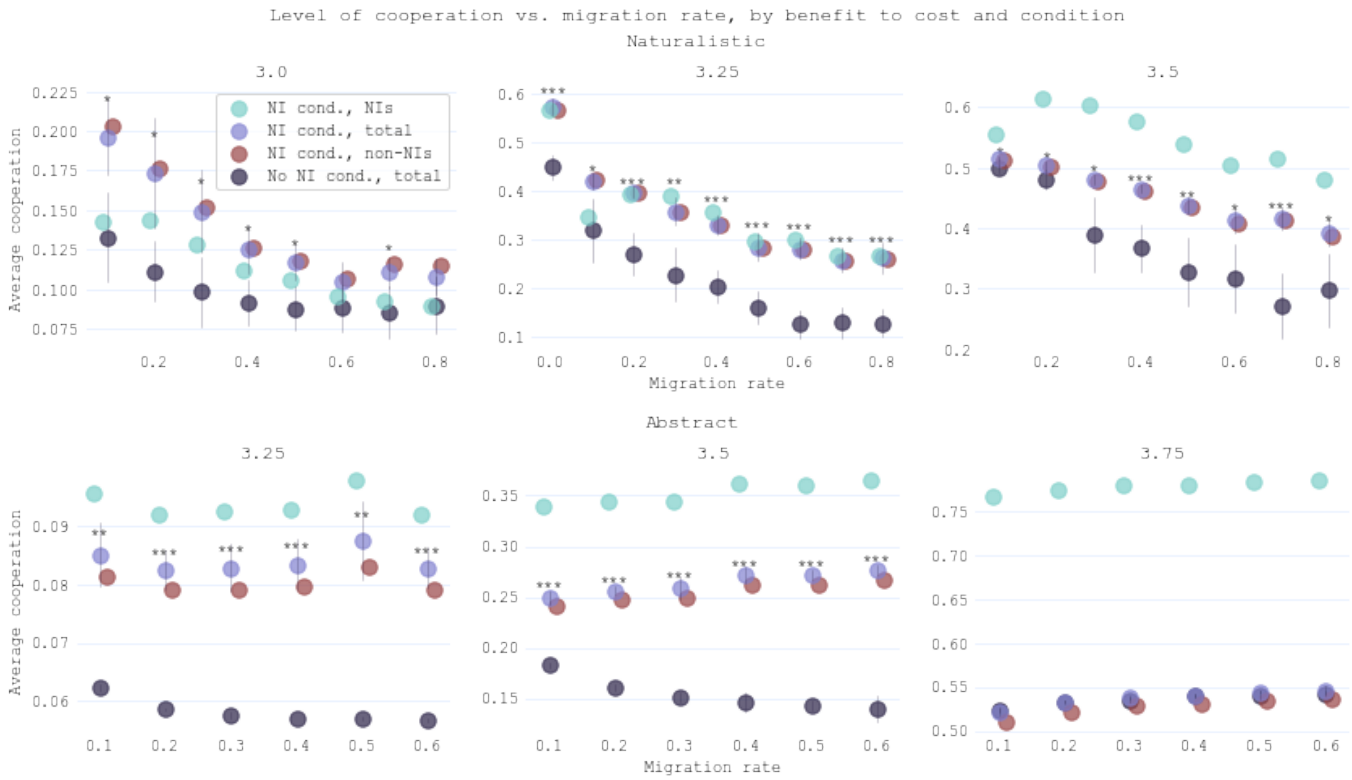

Fig. 2 **Levels of cooperation for various model conditions.** Shown are the mid-range benefit-to-cost ratios (e.g., `3.0`, `3.25`, ...) for which the norm internalizer condition (NI cond.), that is, the condition in which norm internalizers (NIs) were in the strategy space, had significantly higher level of cooperation. For benefit-to-cost ratios above or below the ones shown, the NI condition did not have significantly different cooperation levels.
Some observations to note are **(1)** Whenever the NI condition had significantly different levels of cooperation, it was always higher. **(2)** NIs did not necessarily cooperate more than non-NIs, in fact, in the naturalistic model, NIs made the biggest difference in cooperation levels when they cooperated about equally with everyone else. **(3)** As the benefit-to-cost ratio rises, the level of NI-cooperation rises comparatively to non-NIs. **(4)** The fall-off in cooperation with increasing migration rates is less steep in the NI condition – in fact, in the abstract model, there is no fall-off at all.
For all the following figures, we use the `default` parameter set, $b/c = 3.5$, $p_{\text{mig}} = 0.2$ for the abstract model, and $b/c = 3.25$, $p_{\text{mig}} = 0.5$ for the naturalistic model (these two migration rates end up being equivalent at about a 50% lifetime migration rate); see the appendix for more justification.
The error bars are 95% confidence intervals (t-test), shown only for the total cooperation in each of the two conditions. (*), (**), and (***) respectively indicate the difference between the NI and non-NI condition are different with $p < 0.01$, $p < 0.001$, and $p < 0.0001$ (two-tailed t-test)

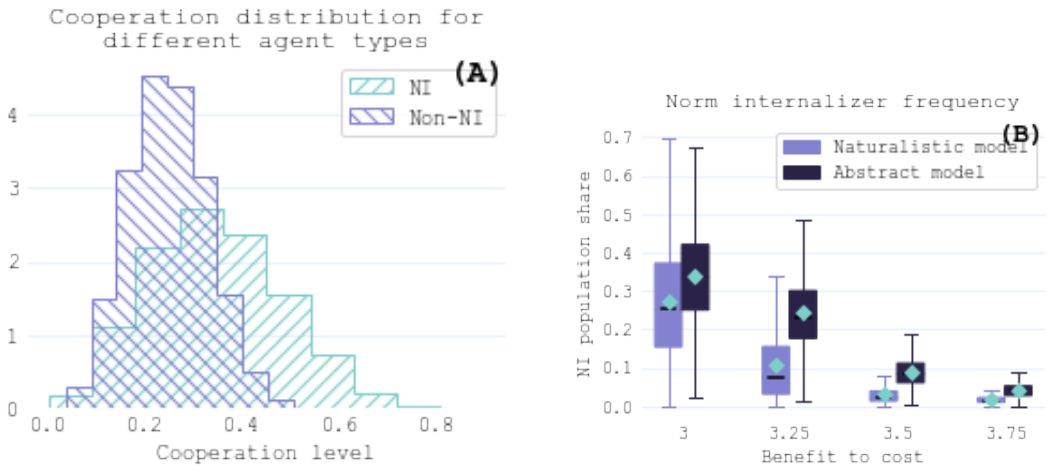

Fig. 3 **Norm internalizer cooperation and population. (A)** Histogram showing the density of cooperation percentages of norm internalizers (NIs) vs non-NIs. NI cooperation tends to be more dispersed in intermediate benefit-to-cost regimes. The naturalistic model exhibits a similar pattern. **(B)** The average population share of NI across rounds, in the naturalistic and abstract models. For reference, NIs made the biggest difference in cooperation at `3.25` in the naturalistic model and `3.5` in the abstract model. The diamonds are the means. The 95% confidence intervals for the median are represented by notches on the box plot, which are so small as to be invisible. Parameters used are the `default` set, except for where b/c is otherwise specified

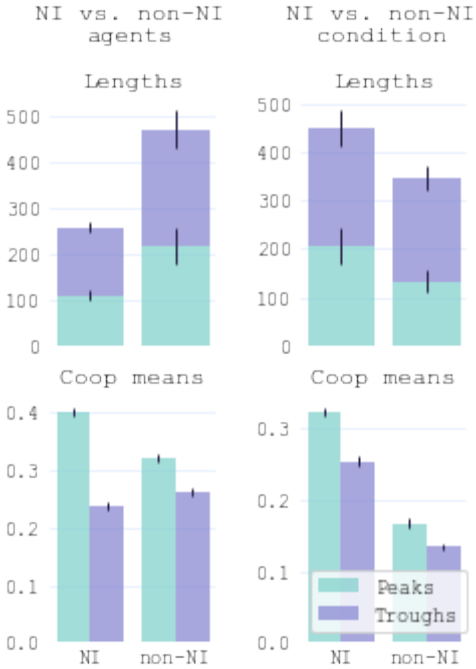

Fig. 4 **Spike-trough comparison for the naturalistic model.** Comparisons of the *lengths* and *cooperation means* of peaks and troughs. Peaks are defined as intervals of at least ten rounds in which cooperation is above average; troughs are the complement of that. At **left**, we compare between cooperation peaks of *norm internalizers* and cooperation peaks of other agent types, all *within the norm internalizer condition*. At **right**, we compare between the NI-condition (NIs are in the strategy space) and the non-NI condition. Though the peaks of norm internalizers are shorter (top left), the peaks in the norm internalizer condition are *longer* (top right). And though the trough cooperation of norm internalizers is slightly lower (bottom left), the trough cooperation levels in the norm internalizer condition are higher (bottom right). Error bars are 99% confidence intervals (z-test), both cooperation means and lengths are significantly different at that level. Parameters are the `default` set.

Table 2 **Key parameters in both models.** Strictly speaking, in the naturalistic model, n and g are the starting number of individuals in a group, and the starting number of groups. Parameters in brackets were only tested in the naturalistic model

Variable	Description	Abstract	Naturalistic
n	number of individuals per group	35	20
g	number of groups	60	10
p_{con}	probability of conflict	1/13	N/A
s	side length of square grid	N/A	10
c_{dist}	cost of foraging on an adjacent square	N/A	5
b/c	benefit to cost ratio	3, 3.25, 3.5, 3.75, 4.0, 4.25	
p_{mig}	probability of migration	0, 0.1, ... 0.6, [0.7, 0.8]	

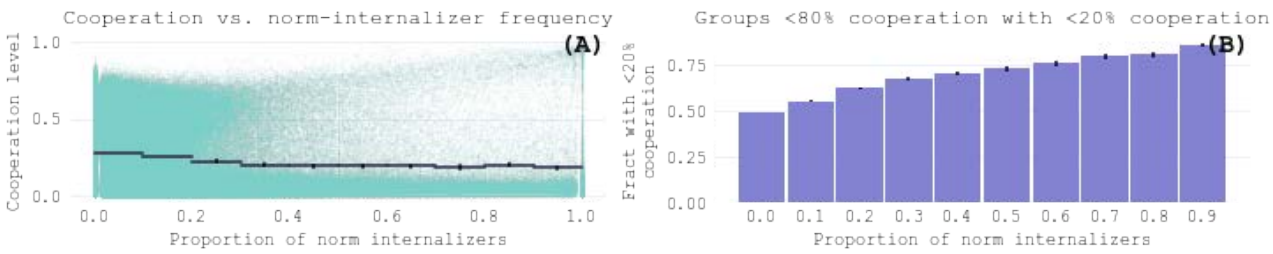

Fig. 5 **Polarization due to norm internalizers. (A)** Scatterplot of cooperation levels of individual groups at difference norm internalizer frequencies in the naturalistic model. The bars are the average cooperation levels for each window. The frequency of norm internalizers does not increase mean cooperation (in fact, there is a mild decrease) but it does polarize cooperation. **(B)** The percentage of groups with cooperation levels below 80% that *also* have cooperation levels under 20%.
Error bars are 99.9% confidence intervals (z-test), parameters are the `default` set.

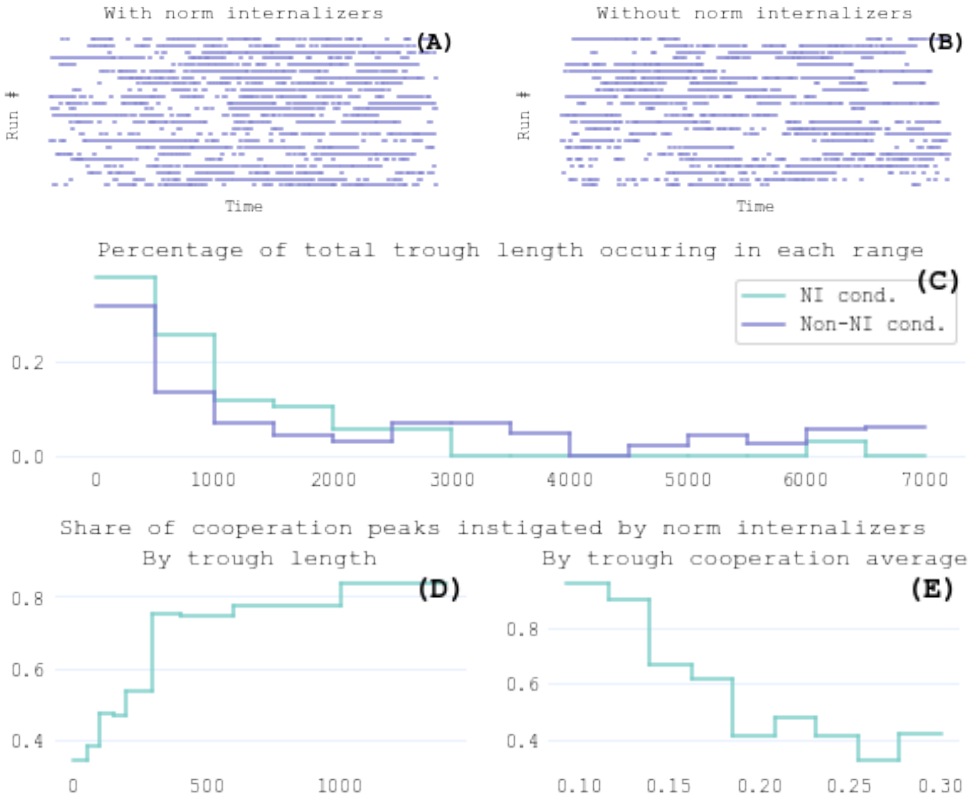

Fig. 6 **Troughs characteristics in the naturalistic model. (A), (B)** The intervals of above-average cooperation (the spikes) are shown as dark lines for each of the simulation runs. While trough lengths for the NI-condition (right) are, on average, slightly longer, it is clear that the non-NI condition (left) has a fatter tail of very long troughs, which can be observed more quantitatively in **(C)**. This histogram plots the probability of being in a trough of various lengths, given that one is in a trough (longer troughs therefore have higher weight since a randomly sampled round is more likely to be in a long than a short trough). **(D), (E)** Plot the share of cooperation spikes *initiated* by norm internalizers after troughs of various lengths and depths, showing that norm internalizers tend to initiate the spikes that follow especially *long* and *low* troughs. Figures use `default` parameters. The abstract model does not exhibit this pattern.

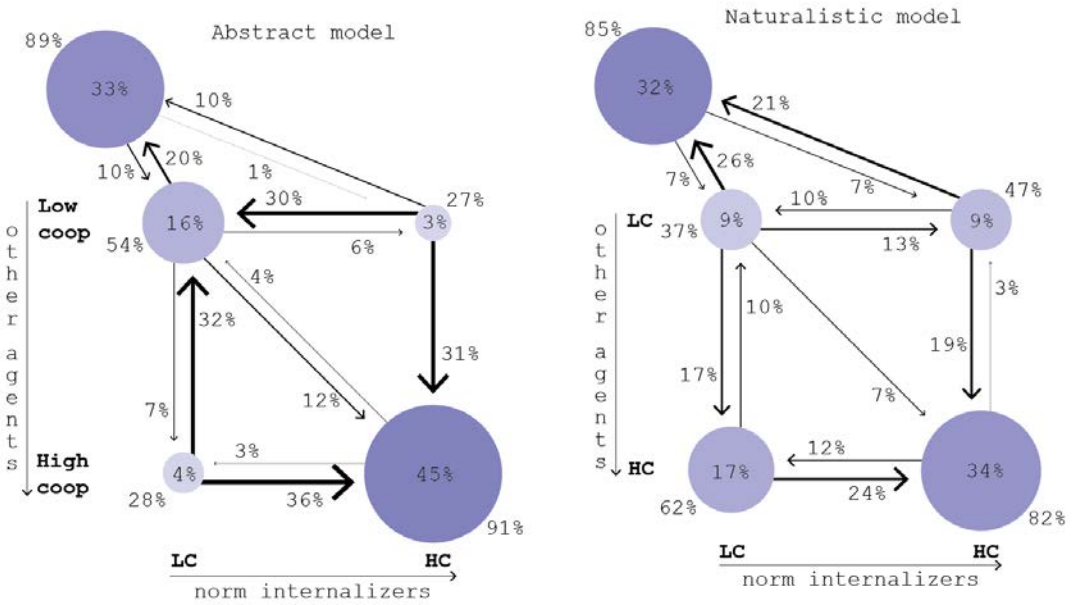

Fig. 7 **Transitions in the naturalistic (left) and abstract (right) models.** Both diagrams represent state transitions in the norm internalizer (NI) condition. The vertical axis shows L(ow)/H(igh) cooperation of non-NIs, the horizontal axis L/H cooperation for NIs. The top left circles represent low cooperation for both, in the special case when the cooperation is more than half a standard deviation below the mean. The arrows shown are the top two outgoing arrows for each node, or all arrows above 5%, whichever set is larger. Circle sizes, along with the percentage listed inside, represent the share of rounds spent in each state; circle tint, along with the percentage listed outside, represents the percentage of self-loops. Each round in the naturalistic (abstract) model is divided into disjoint segments of 25 (10) rounds; transition probabilities are calculated from one window to the next. `Default` parameters were used

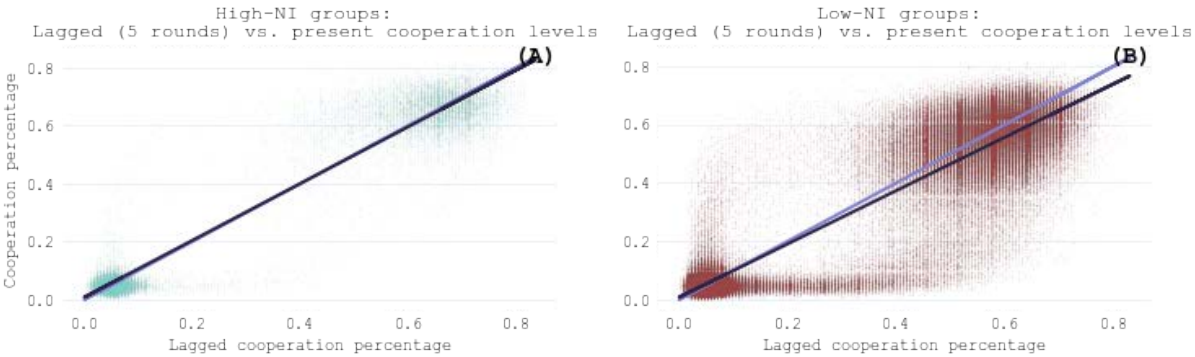

Fig. 8 **Norm internalizers as stabilizers.** Both figures show the level of cooperation in a given round vs. five rounds earlier, for high-NI (>20%) **(A)** and low-NI (<20%)**(B)** groups in the abstract model. The best-fit (dark) is much closer to the slope-1 line (light) in high-NI groups, with coefficients of 0.91 ($\pm$ 0.0015) vs 0.97 ($\pm$ 0.003), respectively (and approximately equal constants). This may not seem like a huge difference, but iterating a cooperation level of 1 through this linear function requires 82 cycles to fall below $\frac{1}{2}$ in high-NI case, while in the low-NI case, it only requires 8 cycles. `Default` parameters were used; confidence intervals are 95% two-tailed t-test. Equivalent figures for the naturalistic model show the same pattern.

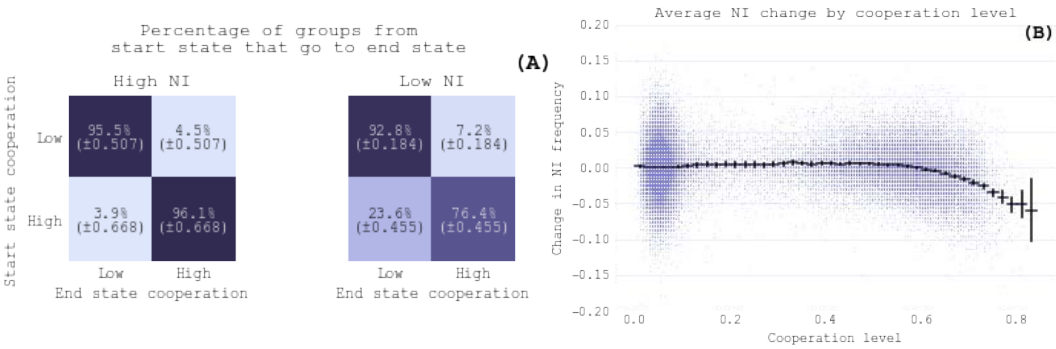

Fig. 9 **Norm internalizers (NIs) as stabilizers, and their shortcoming. (A)** The probability of transitioning from a state of high (>50%) to low (<50%) cooperation for groups that have high (>20%) or low (<20%) proportions of NIs, in the abstract model. High-NI groups maintain high cooperation dramatically better than low-NI groups (though they have a slightly harder time going from low to high cooperation). **(B)** The flaw: here we plot the difference in NI-population at different cooperation levels in the abstract model. Norm internalizer populations tend to fall when cooperation gets high. Both figures show error bars and confidence intervals of 99.9% (z-test), and `default` parameters. Equivalent figures of the naturalistic model show the same pattern.