Evolutionary Human Sciences (2024), **6**, e35, page 1 of 18
doi:10.1017/ehs.2024.37

RESEARCH ARTICLE

Coevolution of norm psychology and cooperation through exapted conformity

Yuta Kido[1,2] and Masanori Takezawa[1,3,4]

[1]Graduate School of Humanities and Human Sciences, Hokkaido University, Sapporo, Japan, [2]Japan Society for the Promotion of Science, Tokyo, Japan, [3]Center for Experimental Research in Social Sciences, Hokkaido University, Sapporo, Japan and [4]Center for Human Nature, Artificial Intelligence and Neuroscience, Hokkaido University, Sapporo, Japan
Corresponding author: Masanori Takezawa; Email: m.takezawa@let.hokudai.ac.jp

(Received 1 February 2024; revised 29 April 2024; accepted 17 July 2024)

Abstract

People willingly follow norms and values, often incurring material costs. This behaviour supposedly stems from evolved norm psychology, contributing to large-scale cooperation among humans. It has been argued that cooperation is influenced by two types of norms: injunctive and descriptive. This study theoretically explores the socialisation of humans under these norms. Our agent-based model simulates scenarios where diverse agents with heterogeneous norm psychologies engage in collective action to maximise their utility functions that capture three motives: gaining material payoff, following injunctive and descriptive norms. Multilevel selective pressure drives the evolution of norm psychology that affects the utility function. Further, we develop a model with exapted conformity, assuming selective advantage for descriptive norm psychology. We show that norm psychology can evolve via cultural group selection. We then identify two normative conditions that favour the evolution of norm psychology, and therefore cooperation: injunctive norms promoting punitive behaviour and descriptive norms. Furthermore, we delineate different characteristics of cooperative societies under these two conditions and explore the potential for a macro transition between them. Together, our results validate the emergence of large-scale cooperative societies through social norms and suggest complementary roles that conformity and punishment play in human prosociality.

Keywords: Evolution; cooperation; social norm; norm psychology; cultural group selection

Social media summary: Norm-psychology can evolve, yielding two distinct states: conformity- and punishment-based cooperative societies.

1. Introduction

Mainstream economic theories start from the fundamental premise that humans pursue self-interest and make decisions based on rational calculations of costs and benefits (Becker, 1976). This assumption is closely related to the logic that the intrinsic gravity of evolution works in the direction of self-interest or the pursuit of survival and reproduction (Dawkins, 1982). However, this self-interested actor model systematically differs from observations: individuals anonymously donate to charity and willing incur personal costs to benefit virtual strangers. Most importantly, humans are the only species in which we observe a large-scale cooperative society among genetically unrelated ephemeral interactants. Interestingly, there is indeed large cultural variability in such cooperation (Gächter et al., 2010; Henrich et al., 2010; Herrmann et al., 2008).

Sociologists view shared societal norms as the root of large-scale cooperation unique to humans. They maintain that social norms, called 'the grammar of society' (Bicchieri, 2005: ix), play a central role in human behaviour by prescribing common value systems within a society (Parsons, 1951). Parsons (1937) argued that internalised norms constitute the self, biasing individuals' behaviours toward values instilled by norms and overpowering egocentric motivations. According to this socialisation theory, cooperation follows certain norms. Likewise, cultural differences manifest because norm-prescribed behaviours differ among societies.

We find a conflict between the economic and sociological views of humankind. In economics, individuals are assumed to be rational and self-regarding, acting to maximise their payoffs, whereas in sociology, they are assumed to be highly socialised agents who prioritise internalised norms over material benefits. From a sociological perspective, it may be possible to explain the unique cooperative behaviour observed in humans. However, this poses a new puzzle: how did humans become socialised beings, deviating from the basic behavioural principle of seeking self-interest?

In this study, we develop a model that incorporates micro-macro dynamics to address this question. At the macro level, we assume two major types of social norms: injunctive and descriptive. At the micro level, that is, at the individual decision-making level, we assume that norm psychology determines one's susceptibility to the influence of social norms. Using agent-based models with evolvable norm psychology, which allow agents to internalise social norms, we explore the possibility and mechanism of coevolution between this socialisation mechanism and large-scale cooperation, as well as the condition for its occurrence.

1.1. Evolution of norm psychology via cultural group selection

Despite the mystery surrounding its evolution, a large body of empirical research implies that our psychologies include a predisposition to follow norms, commonly referred to as 'norm psychology' (Chudek & Henrich, 2011). Children are initially observed to acquire local norms within specific contexts (O'Gorman et al., 2008; Rakoczy et al., 2008), and subsequently experience activation of their brain's reward circuits when they comply with local norms (de Quervain et al., 2004; Rilling et al., 2004). This indicates that the evolved psychological mechanism allows norm compliance to be viewed as a goal rather than a burden. Accordingly, in modelling, some theorists incorporate norm psychology into the utility function (i.e. 'norm-utility models'; see Akçay & van Cleve, 2021; Gavrilets & Richerson, 2017; Gavrilets et al., 2024; Gintis, 2014). Following their lead, our model considers the tradeoff between material utility and social preference derived from norm compliance, with the weight individuals assigned to social preferences varying depending on their norm psychologies.

To explore the evolutionary origin of norm psychology, which could pave the way for a cooperative society, our model was built on the framework of gene–culture coevolutionary accounts (Boyd & Richerson, 1985; Cavalli-Sforza & Feldman, 1981). Some authors explain the evolution of cooperation focusing on cultural processes that homogenise behaviours within groups, followed by selection among groups with large variations (referred to as 'cultural group selection' theory; see Boyd & Richerson, 1985; Henrich, 2004; Smith, 2020). Given high levels of migration, genetic group variations are difficult to sustain; however, cultural learning may allow for homogeneous groups and large cultural variations. If humans with altruistic genetic traits form a cooperative group, group-level selective pressure may outweigh the maladaptive nature of altruism at the individual level. In this framework, Chudek and Henrich (2011) argued that the evolution of norm psychology, which makes us socialise even under altruistic norms, was a crucial step on the path to large-scale cooperation. Following their lead, we develop a coevolutionary model that combines both cultural and genetic evolutionary processes to explain the evolution of cooperation.

1.2. Classification of social norms

Extensive research has been conducted on social norms and cooperation. Here, we focus on specific types of social norms, injunctive and descriptive norms, identified in previous empirical studies as

https://doi.org/10.1017/ehs.2024.37 Published online by Cambridge University Press

important in influencing prosocial behaviour (Cialdini et al., 1990, 1991; Kallgren et al., 2000). Our models were designed to incorporate these norm features.

Injunctive norms are shared standards of behaviour that are expected in a social context. They represent exogenous rules or moral codes, transmitted to the next generation as moral values. Individuals internalising injunctive norms develop social preferences, potentially pursuing virtues benefiting groups at a cost to themselves. Studies have shown that people develop prosocial behaviour according to norms specific to their social groups (House et al., 2013, 2019, 2020; Sutter & Kocher, 2007). Theoretical studies also point to the possibility that cooperation and norm psychology, which internalise injunctive norms, have coevolved in a cooperative dilemma. According to Gavrilets and Richerson (2017), on which our model is based, injunctive norms that promote punishment are likely to be more effective in facilitating coevolutionary processes.

Descriptive norms refer to how common the behaviour is in the social setting. Unlike injunctive norms, descriptive norms depend on individual behaviour. Internalising descriptive norms leads to a preference for following the majority. In other words, it results in a form of social learning process, 'conformity', defined as adopting the most prevalent behaviour (Boyd & Richerson, 1985; Henrich & Boyd, 1998; Whiten et al., 2005). Substantial empirical evidence shows that descriptive norms influence cooperation ('conditional cooperation'; e.g. Fischbacher et al., 2001) and punitive behaviour ('conditional punishment'; Hertz, 2021). However, descriptive norms can also perpetuate detrimental or antisocial behaviours (e.g. smoking, littering, and delinquency) within a group (Schultz et al., 2007). No consistent conclusions have been drawn from theoretical studies on whether or how conformity has coevolved with cooperation (Denton et al., 2020; Efferson et al., 2016; Molleman et al., 2013; Peña et al., 2009; Romano & Balliet, 2017). Theoretical studies have shown that conformity works in tandem with punishment to promote cooperation (Andresguzman et al., 2007; Henrich & Boyd, 2001). However, our models markedly differ from previous models in assuming that a conformist learning strategy is culturally acquired depending on local environments and genetic traits, or norm psychologies. Overall, conformity is at work in real cooperative dilemmas; however, its evolutionary potential in such an environment remains uncertain.

Exaptation of conformity

Conformity (i.e. norm psychology of descriptive norms) was presumably selected to allow us to develop adaptive behaviours beyond cooperation. Mathematical models reveal that conformity is evolutionarily favoured under a wide range of conditions because descriptive norms serve efficiency and accuracy functions, especially in spatially and temporally variable environments (Henrich & Boyd, 1998). Evidence from various species (2-year-old children, Haun et al., 2014; primates, van de Waal et al., 2013; birds, Aplin et al., 2015) supports the idea that conformity can be regarded as a primitive capacity in our psychological mechanism. Hence, it is reasonable to assume that humans were equipped with norm psychology to internalise descriptive norms even before they faced the problem of cooperation. Therefore, just as bird feathers evolved to regulate body temperature and later adapted for flight, conformity probably exapted, evolving in other domains, and then serving one another in the domain of cooperation (Gould & Vrba, 1982). By exapted conformity, we mean a conformity that has evolved to some extent in other domains and has been brought into the domain of cooperation.

Although a large body of literature suggest evolved norm psychologies for different types of social norms underlying humans' unique prosociality, no gene–culture coevolutionary model has addressed the questions how, why and under which conditions they evolve in interaction with each other. Here, we begin by describing our models that consider both cultural and genetic process. We built the models with the following aims, hoping to contribute to the debate about whether humans can evolve from egocentric to social agents. The first aim is to understand the nature of injunctive norms that facilitate norm psychology to coevolve with cooperation, extending the model devised by Gavrilets and Richerson (2017). Second, we explore whether norm psychology for descriptive norms (i.e. conformity) is adaptive in cooperation domains and impactful on the coevolutionary process. Third, we explore the coevolutionary scenario under the assumption of exapted conformity.

https://doi.org/10.1017/ehs.2024.37 Published online by Cambridge University Press

2. Models

We extend the agent-based model (Gavrilets & Richerson, 2017) to simulate gene–culture coevolutionary process in which individuals with heterogeneous norm psychologies engage in collective actions under the influence of two types of social norms. The major parameters in our model are shown in Table 1. This model allows us to explore the possibility that genetic evolution of norm psychology leads to cultural evolution of large-scale cooperation through socialisation of social norms. We assumed a large population of asexual individuals across groups (G), each consisting of 16 members (n). Throughout their lifetimes, group members have opportunities to participate in collective actions for over 40 rounds. The payoff structure of collective actions belongs to a general class of social dilemmas with conflicting interests between individuals and groups (see the supporting information (SI), Text S1.3 for the formulation of the payoff structure, and the SI, Figure S1-1 for the payoff function). Below, we begin by describing the cultural process that involves collective actions and social norms. We then explain the relationship between norm psychology and utility function, which drives an individual's ontogenetic plasticity. Finally, we describe the genetic evolution that optimises fitness.

2.1. Collective actions and social norms

We assume that agents decide whether to participate in two forms of prosocial behaviour in collective action: cooperation (denoted by variable x) and punishment (denoted by variable y). Both are binary strategies that incur a cost for the actor. The payoff π_{CA}, which represents the benefit accruing from the collective action depending on the number of cooperators in the group, is distributed among all group members. Punishers harm all defectors in the group. Then, depending on each agent's strategy (x, y), strategic costs were subtracted from π_{CA}, resulting in the material payoff $\pi(x, y)$ for the individual in each round (see the SI, Text S1.3 for detailed settings regarding the strategy and material payoffs).

Following Gavrilets and Richerson (2017), we assume that these prosocial behaviours are encouraged by the injunctive norm, which is characterised by two non-negative parameters: the normative value of cooperation (denoted by variable v_x) and the normative value of punishment (denoted by variable v_y). We consider the injunctive norm to be exogenously given and constant throughout all generations, but the dynamics of how it is learned and adopted are endogenous to genetic and cultural evolution. This reflects a common situation across human history, in which behaviour is labelled as good or bad but enforcement is not centrally implemented. Furthermore, our models incorporated descriptive norms as typical behaviours within groups. Descriptive norms themselves change

Table 1. Parameters

Symbol	Definition	Value
G	Number of groups	500
n	Group size	16
T	Number of generations	30,000
x	Behavioural trait of cooperation	Variable in {0, 1}
y	Behavioural trait of punishment	Variable in {0, 1}
v_x	A constant determining the strength of injunctive norm for cooperation	{0.0, 0.1, …, 1.0}
v_y	A constant determining the strength of injunctive norm for punishment	{0.0, 0.1, …, 1.0}
$\tilde{X}$	Frequency of cooperators among other group members	Variable in [0.0, 1.0]
$\tilde{Y}$	Frequency of punishers among other group members	Variable in [0.0, 1.0]
α_i	Genetic trait of injunctive norm-psychology	Evolvable in [0.0, 1.0]
α_d	Genetic trait of descriptive norm-psychology	Evolvable in [0.0, 1.0]

https://doi.org/10.1017/ehs.2024.37 Published online by Cambridge University Press

dynamically over a lifetime in the following manner: the frequency of other group members' behaviours (including antisocial ones) in the t_{th} round determines the content and strength of the descriptive norm in the $t+1_{th}$ round.

2.2. Norm psychology and strategy revision

We consider agents that update their strategies with a probability of 0.25 every round throughout their lifetime based on the myopic optimisation algorithm (Sandholm, 2010), which means that individuals choose the optimal strategy based on the others' ones in the previous round, producing the best response dynamics. Individuals revise a combination of strategies (x, y) to maximise the following utility function, considering both material payoffs and social norms:

$$\begin{aligned} \mu_{\alpha_i \alpha_d}(x, y) = \frac{(1-\alpha_i)(1-\alpha_d)}{(1-\alpha_i)(1-\alpha_d)+\alpha_i+\alpha_d} \cdot \pi(x, y) \\ + \frac{\alpha_i}{(1-\alpha_i)(1-\alpha_d)+\alpha_i+\alpha_d} \cdot (v_x \cdot x + v_y \cdot y) \\ + \frac{\alpha_d}{(1-\alpha_i)(1-\alpha_d)+\alpha_i+\alpha_d} \cdot [\tilde{X} \cdot x + (1-\tilde{X})(1-x) + \tilde{Y} \cdot y + (1-\tilde{Y})(1-y)]. \end{aligned} \tag{1}$$

We assume that the influence of social norms is determined by the norm psychology parameter (denoted by $\alpha \in [0.0, 1.0]$). We clearly distinguish between norm psychology of injunctive norm (denoted by α_i) and descriptive norm (denoted by α_d). These genetic traits (α_i and α_d) evolve biologically, allowing for heterogeneously socialised agents and therefore changes in the optimal strategy. The first term in Equation (1) corresponds to the preference for material payoff $\pi(x, y)$ that an agent obtains in each round. Low values of α_i and α_d make agents payoff-oriented because agents place the weight on this preference depending on the value of $(1-\alpha_i)(1-\alpha_d)$. Agents with high value of α_i have a greater weight in the second term. In other words, they obtain higher utility by following the injunctive norm, whose content is characterised by the exogenous parameters v_x and v_y. On the other hand, $\tilde{X}$ and $\tilde{Y}$ are the frequencies of each strategy among other group members, corresponding to the strength of descriptive norms. Altogether, agents with high value of α_d assign more weight to the third term and get higher utility from conforming to others. Note that descriptive norms can encourage selfish behaviours (i.e. $x=0$, $y=0$) in our models when few other group members adopt a prosocial strategy (i.e. low values of $\tilde{X}$ and $\tilde{Y}$), in contrast to injunctive norms, which encourage only prosocial behaviours (see the SI, Text S1.3 for the detailed settings of the strategy revision algorithm).

2.3. Multilevel selection pressure on norm psychology

Each individual was characterised by genetic traits, denoted by α_i and α_d. The initial values are taken from the uniform distribution of $(\alpha_i, \alpha_d) \in [0, 0.05]^2$ with the exception of the exaptation of conformity assumed in the model (higher initial distribution of α_d; Table 2, Model 3). This means that the population starts with self-interested agents with poor socialisation abilities. The evolution of norm psychology is governed by natural selection. Depending on their success in life, multilevel selection drives the evolution of norm psychology. Selection follows a two-level Wright–Fisher process; thus, generations are discrete and nonoverlapping. First, the population in the previous generation is subject to group-level selection, captured by the replication of group j with a probability proportional to the group fitness $w_j = \sum_1^n \sum_1^Q x$, given by the cumulative number of cooperators across 40 rounds (Q) among 16 group members (n). This means that the more benefits a group accumulates, the more likely it is to survive and replicate. Second, the population that survived group-level selection is subject to individual-level selection, captured by the reproduction of individual i with probability proportional to $w_i = w_0 + \bar{\pi}_i$, given by the addition of baseline fitness w_0 and mean of payoffs $\bar{\pi}_i$.

https://doi.org/10.1017/ehs.2024.37 Published online by Cambridge University Press

Table 2. Model assumptions

	Model 1	Model 2	Model 3
Assumed norm psychology	α_i	α_i and α_d	α_i and α_d
Initial distribution of α_i	$\sim U(0.00, 0.05)$	$\sim U(0.00, 0.05)$	$\sim U(0.00, 0.05)$
Initial distribution of α_d		$\sim U(0.00, 0.05)$	$\sim N(0.30, 0.25^2)$

Note: Model 1 is essentially the same as Gavrilets and Richerson's (2017) model. Models 2 and 3 are built upon Model 1 with the addition of the non-exapted and exapted forms, respectively, of α_d.

Under self-interested rationality, individuals would be better off pursuing a material payoff than internalising social norms. Consequently, natural selection at the individual level always decreases reliance on norm psychology. Importantly, however, all agents inhabit a shared environment, which can be influenced by the behaviour of others. If there are many agents around who enforce norms through punishment, defection ($x = 0$) is no longer the optimal strategy; if there are many cooperators around (i.e. large $\tilde{X}$), agents with high α_d potentially obtain the highest utility from cooperation ($x = 1$). In other words, through this coexistence, norm psychology can make groups cooperative, resulting in more success than uncooperative groups, so that norm psychology can be favoured at the group level, although not always (see the SI, Text S1.3 for detailed settings of the multilevel selection algorithm). Finally, half of the group members were randomly selected from each group and migrated to other groups.

3. Results

We consider three models with different settings for social norms. We begin by assuming only injunctive norms and norm psychology α_i, but go on to investigate the impact of descriptive norms and α_d (Table 2). More precisely, first, we replicate the simulation of the model (Gavrilets & Richerson, 2017; Table 2, Model 1) that assumes injunctive norms. However, our simulation differs from previous models in its more fine-grained manipulation of the injunctive norms. Specifically, we exogenously gave normative values for each prosocial behaviour (x, y), varying from 0 to 1 in intervals of 0.1 as (v_x, v_y), yielding 121 simulated combinations of injunctive norms (much more than nine combinations in Gavrilets and Richerson (2017) that have each normative value $v_* \in \{0.0,\ 0.5,\ 1.0\}$). Second, we extended Model 1 by allowing for the influence of descriptive norms and the corresponding norm psychology α_d (Table 2, Model 2). Third, we modelled the exapted conformity by setting an initial distribution of α_d higher as normal distribution $N(0.30, 0.25^2)$ (Table 2, Model 3). Note that the results of Model 3 do not depend, qualitatively, on the specific shape of initial distributions (see the SI, Figure S1, for results under other assumptions about exaptation).

In the analysis, we consider both the steady-state values, approximated by the values in the last generation, and the temporal dynamics. As for behavioural data (i.e. x, y), we report the frequency in the last round. All simulations were routinely run for 30,000 generations to ensure that the genetic traits and resulting behavioural traits reached a steady state as much as possible. In Figure 1, the summary results are illustrated based on the mean value of the last generation for 25 simulation runs.

3.1. Results of Models 1 and 2

For the results of the Model 1 (Figure 1, left column), the third row of the heatmap shows that genetic trait α_i evolves to some extent ($0.2 < \alpha_i < 0.4$) in the top-left region of the parameter space (v_x, v_y). High cooperation rates ($0.8 < x$) and intermediate levels of punishment ($0.3 < y < 0.5$) were observed in the same parameter regions. These observations on the normative conditions for coevolution are broadly consistent with previous findings that cooperation readily evolves under injunctive norms that encourage punishment; in contrast, promoting cooperation is not effective (Gavrilets &

https://doi.org/10.1017/ehs.2024.37 Published online by Cambridge University Press

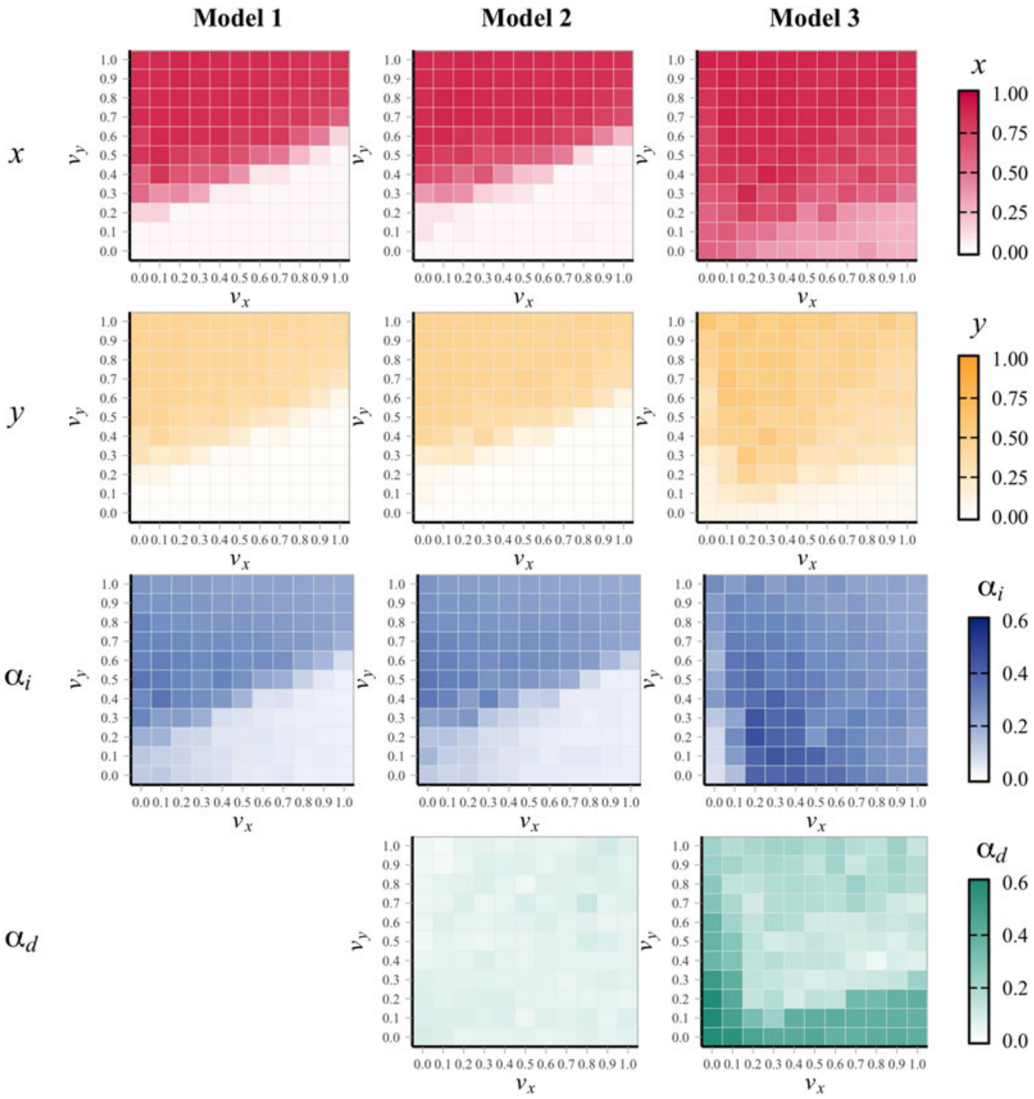

Figure 1. Summary results. Heatmap of x (cooperation), y (punishment), α_i (injunctive norm psychology) and α_d (descriptive norm psychology) for different normative values, v_x (injunctive norm for cooperation) and v_y (injunctive norm for punishment) and three models with different assumptions (Table 2). Shown are averages based on 25 runs for each parameter combination. As for results under other assumptions about exaptation, see the SI, Figure S1.

Richerson, 2017). Furthermore, the increased precision of injunctive norms reveals a new finding: when the norm explicitly values cooperation strongly (i.e. high v_x), α_i tends not to evolve and then, paradoxically, neither does cooperation. In particular, cooperation almost never emerges (mean $x \approx 0.03$) under $(v_x, v_y) = (1.0, 0.5)$. The x, y and α_i values do not markedly differ between Models 1 and 2, while α_d remains very small (Figure 1, middle column). This result suggests that α_d that leads to conformist learning is not favoured in the cooperation domain and, thus, does not influence other evolutionary dynamics and cultural equilibria. In summary, the evolution of cooperation requires an injunctive norm for punishment (v_y) that is sufficiently larger than that for cooperation (v_x) under the non-exapted conformity assumption (Models 1 and 2). In the following subsection, we examine why injunctive norms for punishment are prerequisites for coevolution.

https://doi.org/10.1017/ehs.2024.37 Published online by Cambridge University Press

Evolutionary dynamics (coevolution of α_i and cooperation)

Figure 2a illustrates the evolutionary dynamics typically observed through a representative run (under the setting of $(v_x, v_y) = (0.5, 0.5)$, where cooperation evolved robustly). Here, we draw on the established metric F_{ST} that represents the degree of genotypic differentiation between subpopulations to measure the variation of phenotype of cooperation between groups (red dotted line in Figure 2a). The F_{ST} values range from 0 to 1, with higher values indicating greater differentiation between groups (see the SI, Text S1.3 for the detailed formulation of F_{ST}). For example, when the population is polarised into all-cooperator and all-defector groups, the value of F_{ST} equals 1, whereas it equals 0 when the proportion of cooperators is uniform across all groups. We observed the following dynamics: first, the capacity to internalise the injunctive norm α_i evolves over time (up to the 1200th generation); second, this evolution leads to an increase in cooperative behaviour F_{ST} (1150th–1250th generations). Ultimately, the cooperation rate rises dramatically (up to the 1250th generation). Notably, the genetic F_{ST} (blue dotted line in Figure 2a) remains small while the behavioural F_{ST} is large. These observations are consistent with the evolutionary dynamics of cooperation based on cultural group selection theories (Henrich, 2004) and empirical findings (Bell et al., 2009). Furthermore, the group variations at the three time points (Figure 2b) show the significant effect of punishment on the number of cooperators in the group. In particular, if more than half of the punishers belonged to a group, most group members cooperated. Thus, our results reveal that a fraction of norm enforcers can emerge and shape a cooperative group, which ultimately drives the process in line with the predictions of cultural group selection theory.

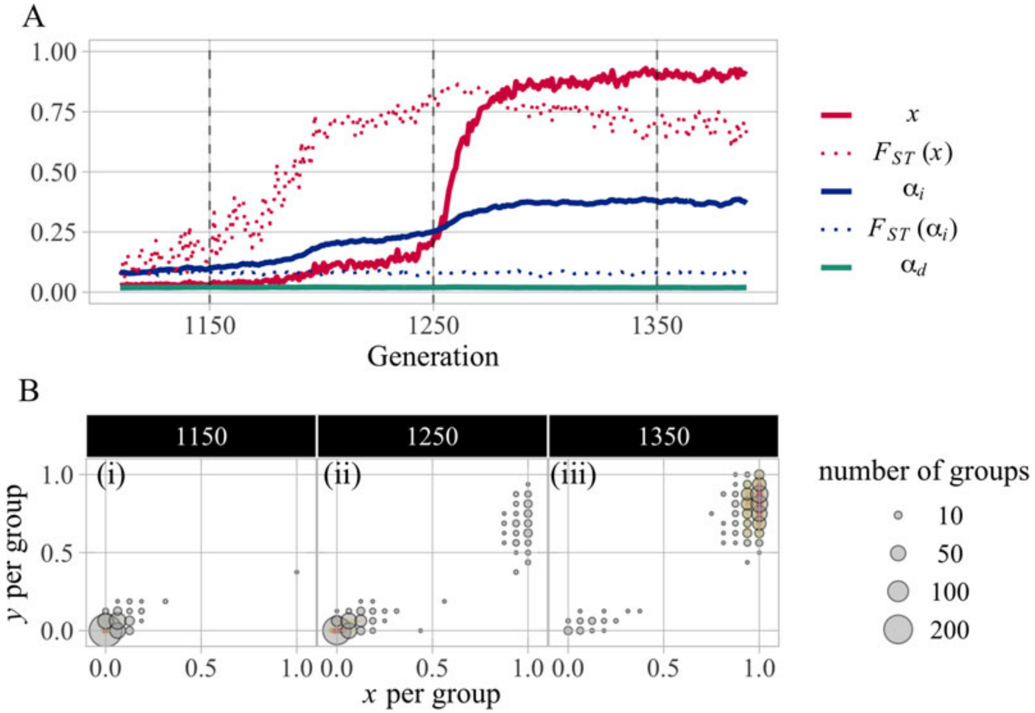

Figure 2. Example of evolutionary dynamics under the setting of non-exapted α_d (Model 2) with $(v_x, v_y) = (0.5, 0.5)$. (a) Mean of α_i (blue solid), genetic F_{ST} of α_i (blue dotted), mean of α_d (green solid), cooperation (red solid) and behavioural F_{ST} of cooperation (red dotted) over the specific generations for a representative simulation. (b) Rate of each behaviour among 500 groups in each generation, with the size representing the number of groups that have the same frequencies of behaviours, x and y. Here, we narrowed down 30,000 to about 300 generations, but afterwards a steady state was reached with some fluctuations (see the SI, Figure S2, for the dynamics over all generations).

https://doi.org/10.1017/ehs.2024.37 Published online by Cambridge University Press

3.2. Result of Model 3

Now, consider the case of exapted conformity (Table 2, Model 3). This time, suppose that simulations start with a population whose initial values of genetic trait α_d are randomly sampled from $N(0.3, 0.25^2)$. Comparisons of the results with Models 2 and 3 explicitly indicate that the latter is more conducive to the coevolution of norm psychology and cooperation. Under the assumption of exapted conformity, cooperation and punishment evolve at high frequencies over a much wider range of conditions than in the other two models. Even when the injunctive norm does not encourage prosocial behaviour at all (i.e. $(v_x, v_y) = (0.0, 0.0)$), intermediate level of cooperation is observed ($x \approx 0.60$, $y \approx 0.12$ and $\alpha_d \approx 0.59$). Moreover, under $(v_x, v_y) = (1.0, 0.5)$, where cooperation did not evolve in Model 1 and Model 2, x becomes greater than .7 with $y \approx 0.32$ and $\alpha_i \approx 0.26$. These results are probably due to the synergistic relationship between exapted conformity and injunctive norms. However, finding a pattern for these effects based on the average of all the simulations (Figure 1) was a challenge. Thus, we classified all simulation results for the last generation based on all parameter values x, y, α_i, and α_d, using clustering analysis by k-means method, which suggests that there are two distinct clusters.

Figure 3 plots the mean value of α_d and the mean frequency of punishment y for each of 25 runs in each injunctive normative condition. Obviously, the two clusters are characterised by a combination of two parameters, α_d and y. In the first cluster (hereafter, Cluster 1), α_d evolves to a surprisingly small value (mean $\alpha_d \approx 0.14$) with a certain number of punishers (mean $y \approx 0.48$) among the population; in the second cluster (hereafter, Cluster 2), exapted α_d remains high or evolves to a higher value (mean $\alpha_d \approx 0.43$) with few punishers (mean $y \approx 0.10$). Cooperation is observed in both clusters, although its extent and mechanisms differ.

To better understand the nature of the two stable states, we identified the simulations closest to the centroid of each cluster (represented by the black points in Figure 3a), and presented the frequency of cooperators and punishers per group in the final generation (Figure 3b, c). Cluster 2, in which conformists are not driven out, is common when either the injunctive norm does not strongly promote punishment (low values of v_y) or does not encourage cooperation ($v_x = 0.0$) (see the SI, Figure, S10b for the relationship between normative values and the cluster ratio in Model 3). In these instances, intermediate levels of cooperation with large group variation are achieved by exapted conformity alone, where low levels of punishment are observed because descriptive norms of punishment are rarely formed and maintained, owing to a net negative cost of punishment. On the other hand, in broader conditions, the population settles down into Cluster 1, characterised by higher y and lower α_d. Figure 3b shows a clear trend in which cooperation is achieved with the punishers. The societies tend to achieve higher cooperation when supported by punishment than by conformity ($\bar{x} \approx 0.86$ in the Cluster 1, while $\bar{x} \approx 0.39$ in the Cluster 2 on average). Moreover, the assumption of exapted conformity expands the basin of attractions for punishment-based cooperative societies (Cluster 1), as well as conformity-based societies (Cluster 2). Then, we explore the dynamics behind the macroscopic change that results in the evolution of cooperation over a wider range of conditions by scrutinising a simulation in the normative condition $(v_x, v_y) = (1.0, 0.5)$, in which cooperation evolves robustly only in Model 3.

Evolutionary dynamics (coevolution of α_i, α_d and cooperation)

Here, through typical temporal dynamics, we show that the mechanism underlying cooperative societies shifts from conformity (descriptive norm) to punishment (injunctive norm). Figure 4 illustrates the dynamics over approximately 3000 generations in an exemplary simulation under the setting of $(v_x, v_y) = (1.0, 0.5)$. In Figure 4a, we observe a series of dynamics consisting of the following three phases: First, a cooperative state with high α_d and low y is achieved (up until the 300th generation). In this phase, the maintenance of cooperation depends on conformist learning, causing significant cultural differences in cooperation between groups, even without punishment (Figure 4b(i)). However, this state, which can be classified as a conformity-based cooperative society (Cluster 2), does not

https://doi.org/10.1017/ehs.2024.37 Published online by Cambridge University Press

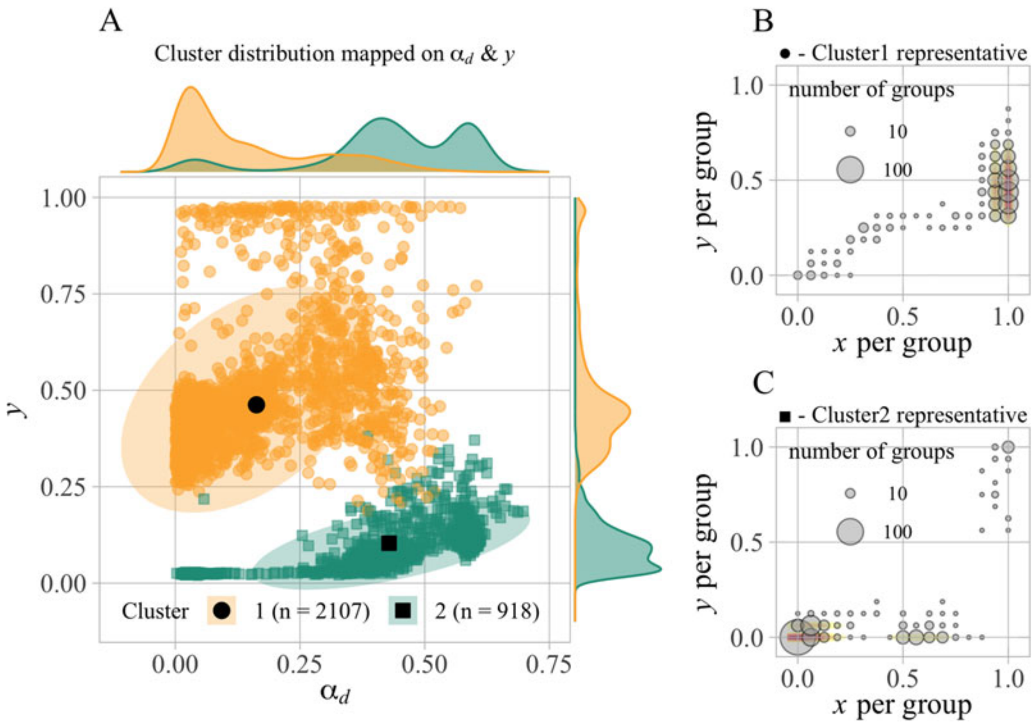

Figure 3. Clustering of all results under the setting of exapted α_d (Model 3). (a) Scatterplot of all simulation results for the last generation, with the mean value of α_d and the frequency of y on the axes, clustered by the k-means method. Results are plotted as yellow circles for Cluster 1 and as green squares for Cluster 2. Ellipses cover about 80% of simulations in each cluster, assuming a multivariate normal distribution. The simulations closest to the centroid of each cluster are shown in the black circle (Cluster 1) and square (Cluster 2). (b, c) Frequency of x and y per group in the representative simulation (i.e. the centroid of each cluster) with the size showing the number of groups whose frequencies of behaviours were the same. As for the analysis of optimal number of clusters and clustering results for Model 2, see the SI, Figure S8.

last long. Instead, a transition occurs in the social state. This is the second phase of the dynamics. During this phase, the mechanism for maintaining cooperation shifts from conformity to punishment, with a temporary decline in cooperation to approximately 50% (around the 1000th–2000th generation). This is illustrated in Fig. 2b(ii), where punishment-based cooperative groups begin to emerge and all agents adopt both prosocial strategies (i.e. $x = y = 1$). Over the course of time, α_i and prosocial behaviours increase considerably at a certain tipping point (around the 2000th generation). Finally, a cooperative society relying on high α_i and y emerges, with very low values of α_d (around the 2500th generation). In Figure 2b(iii), almost all the groups converge to a state with high x, y.

Here, we examine the generality of the phase transition dynamics from conformity-based to punishment-based cooperation observed in Figure 4. Figure 5 plots the 2D state transitions of 25 runs in α_d and y space at 5 time points for three norm value (v_x, v_y) combinations. In Model 2, cooperation only evolved when $(v_x, v_y) = (0.5, 0.5)$, all of which converge to the punishment-based state. In Model 3, cooperation also evolved when $(v_x, v_y) = (0.5, 0.5)$, but pathways to cooperative states markedly differ. In early generations all runs are in the state of cooperation by conformity, and over generations they transition to cooperation by punishment in the upper left of the pane. Moreover, while no runs showed the evolution of cooperation in Model 2 for $(v_x, v_y) = (1.0, 0.5)$, most of the runs in Model 3 exhibited the similar trajectory leading to cooperation.

We further categorised run states into ‘defection’ ($x < 0.5$), ‘cooperation by punishment’ ($x \geq 0.5$ *and* $y \geq \alpha_d$), and ‘cooperation by conformity’ ($x \geq 0.5$ & $\alpha_d > y$), and illustrated their frequency changes over full generations (Figure 6). In Model 2, punishment-based cooperation gradually

https://doi.org/10.1017/ehs.2024.37 Published online by Cambridge University Press

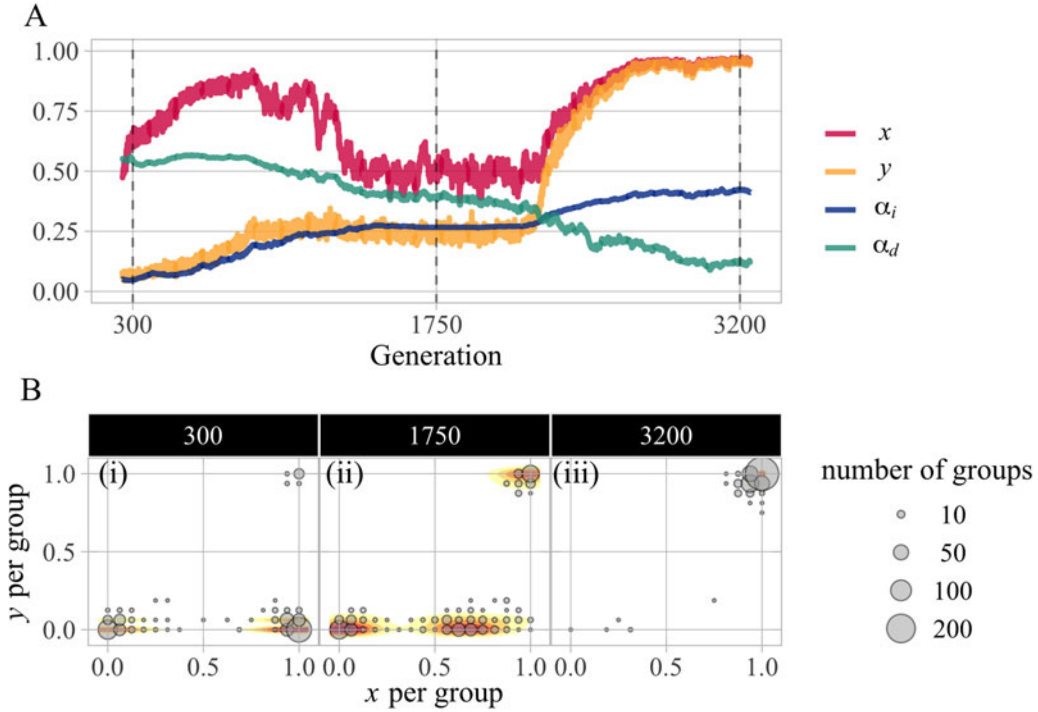

Figure 4. Example of evolutionary dynamics under the setting of exapted α_d (Model 3) with $(v_x, v_y) = (1.0, 0.5)$. (a) Mean of α_i (blue), α_d (green), x (red), and y (yellow) over the specific generations for a simulation. (b) Rate of each behaviour among 500 groups in each generation, with the size representing the number of groups whose frequencies of behaviours were the same. Here, we narrowed down 30,000 to about 3000 generations. Note that thereafter the frequency of punishers decreases by about half, and the mean value of α_i also decreases slightly, reaching a steady state (see the SI, Figure S5, for the dynamics over all generations).

emerged when v_y was sufficiently higher than v_x. In contrast, Model 3 showed initial cooperation by conformity, often following cooperation by punishment. This transitional dynamics led to two notable changes: punishment-based cooperation emerged over broader normative conditions, and it did more rapidly.

4. Discussion

We developed a set of gene–culture coevolutionary models that explore the coevolutionary process of norm psychology and cooperation. Our results confirmed the possibility that a large-scale cooperative society can emerge via norm internalisation with altruistic norms as well as the dynamics underlying coevolution, which is consistent with cultural group selection theories. The evolution of norm psychology can lead to substantial variation between groups, resulting in a large-scale cooperative society. Furthermore, our models allowed us to identify the types of social norms that contribute to the evolution of cooperation through their embodiment by socialised agents, and draw novel connections between different types of social norms and cooperation. In this final section, we highlight the key findings about each social norm and discuss the implications of the results, limitations, and directions for future research.

4.1. Summary

Injunctive norm

Our study provides insight into the conditions for injunctive norms that favour the coevolution of norm psychology and cooperation: sufficiently encouraging punishment compared with cooperation.

https://doi.org/10.1017/ehs.2024.37 Published online by Cambridge University Press

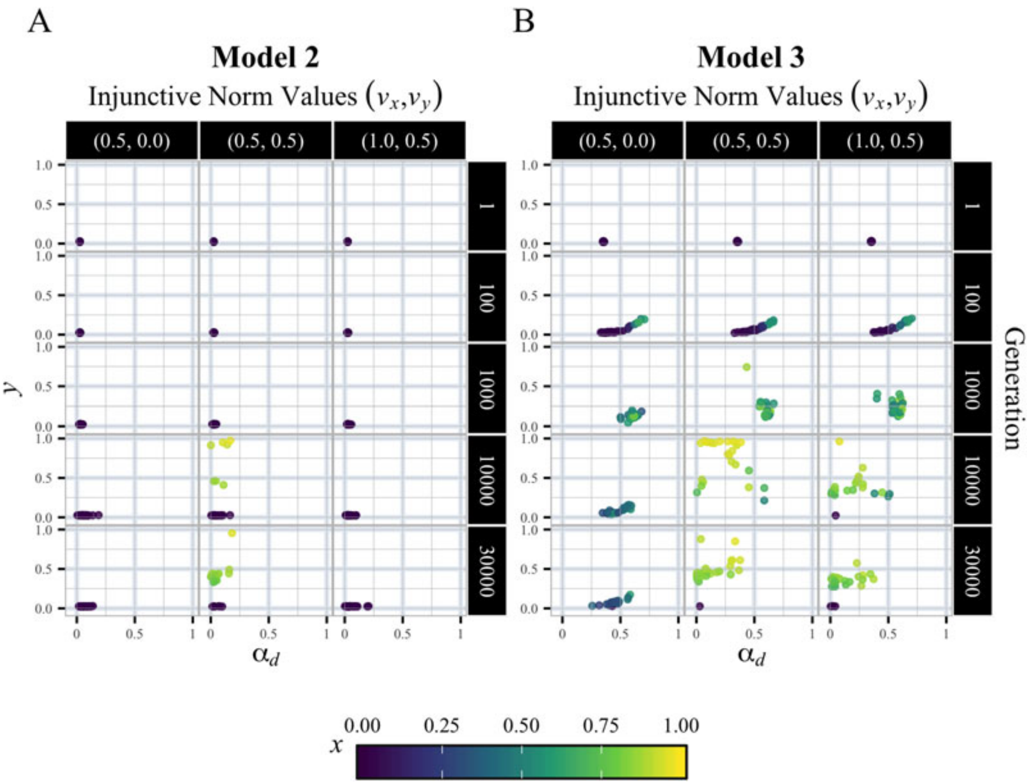

Figure 5. Comparison of temporal dynamics between Models 2 and 3. (a, b) Trajectory of 25 runs at 5 time points (1, 100, 1000, 10,000 and 30,000th generation) in the 2D space of α_d and y, for three combinations of injunctive norm values $(v_x, v_y) = (0.5, 0.0)$, (0.5, 0.5), (1.0, 0.5) in Models 2 and 3. Colour represents cooperation rate at each time point.

This aligns with prior theoretical studies (Gavrilets & Richerson, 2017), highlighting the equilibrium function of punishment that makes any behaviour viable (Boyd & Richerson, 1992). Our analysis reveals how punishment maintains cultural equilibria of cooperation. In a cooperative society underpinned by punishment, individuals are split into two types: vigilantes, who have strongly internalised the punishment norm, and selfish agents, who have only payoff-oriented considerations in decision-making (Figure 3b; see the SI, Figures S3, S4, S6 and S7 for dimorphic populations in genotype $\alpha_i$and phenotypes (x, y)). In groups with vigilantes, the payoff from defection is less than that from cooperation. Consequently, a uniform cooperative group consisting of vigilantes and conditional cooperators emerge to evade punishment.

More interestingly, the comprehensive manipulation of injunctive norms refined the sufficient conditions for the evolution of cooperation. Strongly encouraging cooperation with injunctive norms tends not to favour the norm psychology and eventual cooperation. This suggests that those internalising cooperative norm may face exploitation by free riders, highlighting the potential drawback of injunctive norms in promoting cooperation – a novel finding in our study.

Descriptive norm

Our models also examined the adaptive value of the psychology of internalising descriptive norms (i.e. conformity) in a social dilemma. We demonstrated that conformity is unlikely to evolve from scratch in this domain. Then, we presumed that this domain-general learning capacity was brought into the specific domain of cooperation and work. We showed that this could coevolve with cooperation through the selective force of intergroup competition under this presumption.

https://doi.org/10.1017/ehs.2024.37 Published online by Cambridge University Press

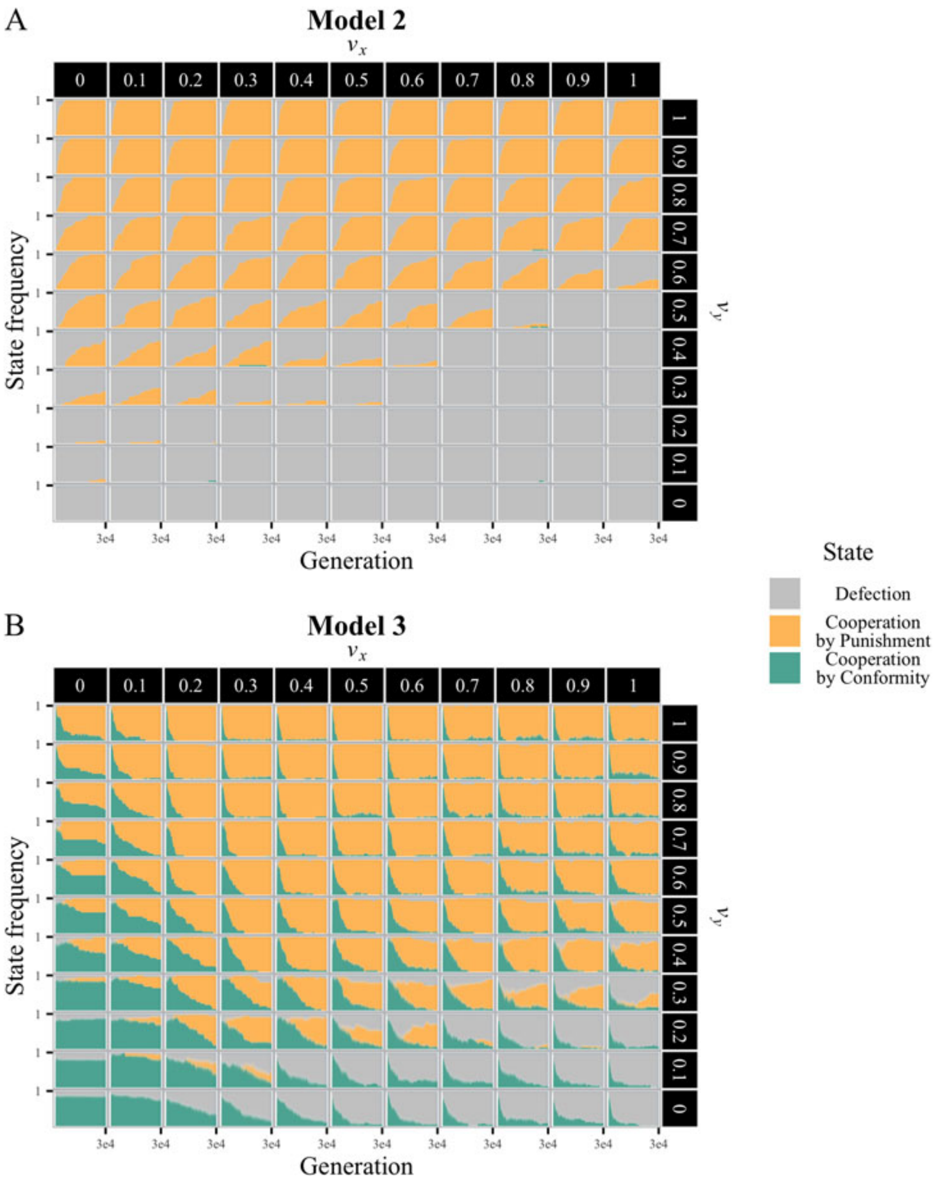

Figure 6. Comparison of state transition dynamics between Models 2 and 3. (a, b) Time series of frequencies of the following three states across all injunctive norm value combinations (v_x, v_y). 'Defection' (grey) is defined as $x < 0.5$, 'cooperation by punishment' (yellow) as $x \geq 0.5$ *and* $y \geq \alpha_d$, and 'cooperation by conformity' (green) as $x \geq 0.5$, $\alpha_d > y$.

However, cooperative societies built upon conformity exhibit different stability features compared with those supported by punishment. Offline simulations, where cooperative groups from online simulations engage in 40 rounds of public goods games again, reveal that these societies are not inherently stable (see the SI, Figure S11, for the online simulation results closest to the centroid of each cluster and the offline simulation results). This instability arises because the emergence and sustainability of cooperative groups hinge on the prevalence of cooperation, and descriptive norms can fluctuate, introducing structural challenges in maintaining cooperation.

https://doi.org/10.1017/ehs.2024.37 Published online by Cambridge University Press

Interplay between Injunctive and Descriptive norm

In the earlier discussions, we highlighted two potent evolutionary drivers of cooperation: punishment and conformity. Moreover, we argue that exapted conformity might have served as a scaffolding for the evolution of punishment, primarily owing to the differing strength of attraction between punishment- and conformity-based cooperative societies and the great speed of cultural relative to genetic evolution. This dynamic process, which leads to the expansion of the basin of attraction for punishment-based cooperation, unfolds as follows. First, conformist transmission initiates a cultural process that establishes and sustains group boundaries. In a mixed population of cooperative and non-cooperative groups, injunctive norm psychology α_i, driving agents to internalise altruistic punishment norms, is likely to be favoured. Thus, with exapted conformity, cooperation evolves under broader normative conditions. In other words, the exaptation assumption can provide a different starting point for the fitness landscape, thereby reaching a higher peak.

This finding partially supports the argument that punishment and conformity played complementary roles in the evolution of prosociality (Henrich & Boyd, 1998, 2001; Andresguzman et al., 2007). However, it does not align with the prediction that two micro mechanisms would work and evolve together in social dilemmas, either culturally (Henrich & Boyd, 2001) or genetically (Andresguzman et al., 2007). Whether the model allows for both cultural and genetic evolution may account for the inconsistency between such arguments and the findings of this study. Our results suggest that, given the assumption that agents acquire learning biases genetically and behaviours culturally, the coexistence of conformity and punishment is unlikely or short-lived. Instead, as elucidated above, each fosters distinct forms of cooperative societies, with macro transitions between them.

4.2. Implications

Our theoretical predictions align with existing empirical evidence, emphasising the pivotal role of the punishment norm in human cooperation. Human proclivity for third-party punishment in response to norm violations has been well documented (Fehr & Fischbacher, 2004; Henrich et al., 2006; Mathew et al., 2013). This study, which demonstrated the spontaneous emergence of vigilantes who punish willingly at a cost, provides an explanation for those dispositions. As theoretically suggested (Akçay & van Cleve, 2021), the internalisation of external punishments by individuals could form a more stable foundation for social order based on punishment, potentially evolving into formalised institutions like law enforcement (North, 2010).

Conformity, according to our predictions, plays a crucial role in the evolutionary process of cooperation, particularly in environments with high migration rate parameter $m = 0.5$, which exceeds the observed migration rates among actual hunter–gatherer populations (Marlowe, 2005). Theoretically, the increase in migration reduces not only genetic but also cultural differences among groups, thus making the condition for the evolution of cooperation more stringent. However, in our models, conformity mitigates the condition (see the SI, Figures S15 and S16, for a summary graph with smaller and larger migration rates, m). Nonetheless, conformity by itself lacks the ability to establish a robust order and is susceptible to negative selection over time. This aligns with the theoretical preference for ‘weak conformity’ in previous studies (Claidière et al., 2012; Kandler & Laland, 2009), which has empirical support (Eriksson & Coultas, 2009; McElreath et al., 2005). Taken together, the two proximate mechanisms maintaining social norms and resulting normative regularities provide clues as to the framework in exploring the potential of non-human norms and interpreting empirical data (Andrews et al., 2024).

4.3. Limitations and future directions

However, the conclusions drawn from our simulations warrant caution owing to some impactful assumptions on results. A key assumption involves intergenerational strategy transmission, where we conservatively posit random strategy acquisition at the beginning of each generation. If,

https://doi.org/10.1017/ehs.2024.37 Published online by Cambridge University Press

alternatively, we assume vertical transmission from parents, the dominance of conformity in maintaining uniformity proves too strong for cooperation to emerge once defection stabilises (see the SI, Figure S12, for a summary graph with a vertical transmission setting). This finding is consistent with previous studies asserting that strong conformity can impede the spread of adaptive variants (Kandler & Laland, 2009).

Moreover, our model relies on several assumptions regarding social norms. First, we represented injunctive and descriptive norms as independent, following Cialdini et al. (1990), although they can be viewed as almost identical or strongly related. However, real-world observations indicate that descriptive and injunctive norms can sometimes be incongruent (Cialdini et al., 1991; Ewing, 2001; Perkins & Berkowitz, 1986). Although this assumption led to novel findings, it concurrently introduces a limitation: the model does not consider the endogeneity of injunctive norms, avoiding potential confounding effects arising from two endogenous norms. Of course, some argue in favour of this assumption, positing that, in pre-modern societies, norms were external rules not generated within the society, implying that injunctive norms were not subject to endogenous evolution (Giddens, 1991). However, historical events, such as the Reformation, highlight conflicts between societies with different injunctive norms significantly affecting behaviours and beliefs. Future studies are essential to explore the selection process among societies with endogenous and evolving injunctive norms shaped through continuous social interactions. Secondly, for simplicity, we assumed that the norm psychologies underlying social norms were invariant within generations. However, empirical evidence suggests systematic interplay between two types of norm psychologies over cultural time (Bicchieri, 2005; Bicchieri & Xiao, 2009). The observed transition from descriptive to injunctive norms over evolutionary timescales in this study may potentially occur in a developmental process (Heyes, 2023). This will contribute to a more nuanced understanding of how different types of social norms coexist and influence human behaviours.

4.4. Conclusion

This study explores the prospect of socialisation by humans even under altruistic norms. Sociologists argue that behind large-scale human cooperation lie norms that embody common values and restrain self-interested behaviour, treating agents as social actors shaped by norm internalisation. However, criticisms of the teleological nature of the 'over-socialised' concept have prompted a deeper exploration of the functional significance of internalising social norms. Thus, this study bridges sociology, economics and biology to scrutinise the validity of socialisation theory. Addressing the initial question posed in this study, it is now conceivable that humans can indeed be socialised into prosocial norms. This study yields two key insights. First, injunctive norms that prioritise punishment over cooperation prompt internalisation, fostering the evolution of cooperation. Second, the psychological mechanism of internalising descriptive norms may establish the prerequisites for a large, cooperative society sustained by punishment. These findings contribute to a multidisciplinary understanding of human social dynamics, shedding light on the nuanced interplay between individual psychology, social norms and cooperative behaviour.

Supplementary material. To view supplementary material for this article, please visit https://doi.org/10.1017/ehs.2024.37

Acknowledgements. We would like to thank Editage (www.editage.com) for English language editing.

Author contributions. Y.K. and M.T. designed the computational models and computer simulations; Y.K. conducted computer simulations and data analysis; and Y.K. and M.T. wrote the manuscript.

Financial support. This work was supported by JSPS Research Fellow (Y.K., grant number 22J20339), Grant Program of Graduate School of Humanities and Human Sciences, Hokkaido University (awarded to Y.K.), the foundation for the Fusion of Science and Technology (awarded to Y.K.), and JSPS KAKENHI Grant Number JP17KK0060 and JP17H02621 awarded to M.T.

Conflicts of interest. The authors declare no conflict of interest existing.

https://doi.org/10.1017/ehs.2024.37 Published online by Cambridge University Press

Research transparency and reproducibility. All simulation codes are available at https://github.com/kidd000/Kido_Takezawa2024.

Data availability statement. n/a.

References

Akçay, E., & van Cleve, J. (2021). Internalizing cooperative norms in group-structured populations. In Wilczynski, W. & Brosnan, S. (Eds.), *Cooperation and conflict: The interaction of opposites in shaping social behavior* (pp. 26–44). Cambridge University Press.

Andresguzman, R., Rodriguezsickert, C., & Rowthorn, R. (2007). When in Rome, do as the Romans do: The coevolution of altruistic punishment, conformist learning, and cooperation. *Evolution and Human Behavior: Official Journal of the Human Behavior and Evolution Society*, *28*(2), 112–117.

Andrews, K., Fitzpatrick, S., & Westra, E. (2024). Human and nonhuman norms: A dimensional framework. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *379*(1897), 20230026.

Aplin, L. M., Farine, D. R., Morand-Ferron, J., Cockburn, A., Thornton, A., & Sheldon, B. C. (2015). Experimentally induced innovations lead to persistent culture via conformity in wild birds. *Nature*, *518*(7540), 538–541.

Becker, G. S. (1976). *The economic approach to human behavior*. University of Chicago Press.

Bell, A. V., Richerson, P. J., & McElreath, R. (2009). Culture rather than genes provides greater scope for the evolution of large-scale human prosociality. *Proceedings of the National Academy of Sciences*, *106*(42), 17671-17674.

Bicchieri, C. (2005). *The grammar of society: The nature and dynamics of social norms*. Cambridge University Press.

Bicchieri, C., & Xiao, E. (2009). Do the right thing: But only if others do so. *Journal of Behavioral Decision Making*, *22*(2), 191–208.

Boyd, R., & Richerson, P. J. (1985). *Culture and the evolutionary process*. University of Chicago Press.

Boyd, R., & Richerson, P. J. (1992). Punishment allows the evolution of cooperation (or anything else) in sizable groups. *Ethology and Sociobiology*, *13*(3), 171–195.

Cavalli-Sforza, L. L., & Feldman, M. W. (1981). *Cultural transmission and evolution: A quantitative approach*. Princeton University Press.

Chudek, M., & Henrich, J. (2011). Culture–gene coevolution, norm-psychology and the emergence of human prosociality. *Trends in Cognitive Sciences*, *15*(5), 218–226.

Cialdini, R. B., Reno, R. R., & Kallgren, C. A. (1990). A focus theory of normative conduct: Recycling the concept of norms to reduce littering in public places. *Journal of Personality and Social Psychology*, *58*(6), 1015–1026.

Cialdini, R. B., Kallgren, C. A., & Reno, R. R. (1991). A focus theory of normative conduct: A theoretical refinement and reevaluation of the role of norms in human behavior. In M. P. Zanna (Ed.), *Advances in experimental social psychology*, *24*, 201–234. Academic Press.

Claidière, N., Bowler, M., & Whiten, A. (2012). Evidence for weak or linear conformity but not for hyper-conformity in an everyday social learning context. *PloS One*, *7*(2), e30970.

Dawkins, R. (1982). *The extended phenotype: The gene as the unit of selection*. Freeman.

Denton, K. K., Ram, Y., Liberman, U., & Feldman, M. W. (2020). Cultural evolution of conformity and anticonformity. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(24), 13603–13614.

de Quervain, D. J.-F., Fischbacher, U., Treyer, V., Schellhammer, M., Schnyder, U., Buck, A., & Fehr, E. (2004). The neural basis of altruistic punishment. *Science*, *305*(5688), 1254–1258.

Efferson, C., Lalive, R., Cacault, M. P., & Kistler, D. (2016). The evolution of facultative conformity based on similarity. *PloS One*, *11*(12), e0168551.

Eriksson, K., & Coultas, J. C. (2009). Are people really conformist-biased? An empirical test and a new mathematical model. *Journal of Evolutionary Psychology*, *7*(1), 5–21.

Ewing, G. (2001). Altruistic, egoistic, and normative effects on curbside recycling. *Environment and Behavior*, *33*(6), 733–764.

Fehr, E., & Fischbacher, U. (2004). Third-party punishment and social norms. *Evolution and Human Behavior: Official Journal of the Human Behavior and Evolution Society*, *25*(2), 63–87.

Fischbacher, U., Gächter, S., & Fehr, E. (2001). Are people conditionally cooperative? Evidence from a public goods experiment. *Economics Letters*, *71*(3), 397–404.

Gächter, S., Herrmann, B., & Thöni, C. (2010). Culture and cooperation. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *365*(1553), 2651–2661.

Gavrilets, S., & Richerson, P. J. (2017). Collective action and the evolution of social norm internalization. *Proceedings of the National Academy of Sciences*, *114*(23), 6068–6073.

Gavrilets, S., Tverskoi, D., & Sánchez, A. (2024). Modelling social norms: An integration of the norm-utility approach with beliefs dynamics. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *379*(1897), 20230027.

Giddens, A. (1991). *Modernity and self-identity: Self and society in the late modern age*. Stanford University Press.

https://doi.org/10.1017/ehs.2024.37 Published online by Cambridge University Press

Gintis, H. (2014). *The bounds of reason: Game theory and the unification of the behavioral sciences* (revised ed.). Princeton University Press.

Gould, S. J., & Vrba, E. S. (1982). Exaptation – A missing term in the science of form. *Paleobiology*, *8*(1), 4–15.

Haun, D. B. M., Rekers, Y., & Tomasello, M. (2014). Children conform to the behavior of peers; Other great apes stick with what they know. *Psychological Science*, *25*(12), 2160–2167.

Henrich, J., (2004). Cultural group selection, coevolutionary processes and large-scale cooperation. *Journal of Economic Behavior & Organization*, *53*(1), 3–35.

Henrich, J., & Boyd, R. (1998). The evolution of conformist transmission and the emergence of between-group differences. *Evolution and Human Behavior: Official Journal of the Human Behavior and Evolution Society*, *19*(4), 215–241.

Henrich, J., & Boyd, R. (2001). Why people punish defectors. Weak conformist transmission can stabilize costly enforcement of norms in cooperative dilemmas. *Journal of Theoretical Biology*, *208*(1), 79–89.

Henrich, J., McElreath, R., Barr, A., Ensminger, J., Barrett, C., Bolyanatz, A., … Ziker, J. (2006). Costly punishment across human societies. *Science*, *312*(5781), 1767–1770.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *The Behavioral and Brain Sciences*, *33*(2–3), 61–83; discussion 83-135.

Herrmann, B., Thöni, C., & Gächter, S. (2008). Antisocial punishment across societies. *Science*, *319*(5868), 1362–1367.

Hertz, U. (2021). Learning how to behave: cognitive learning processes account for asymmetries in adaptation to social norms. *Proceedings of the Royal Society B: Biological Sciences*, *288*(1952), 20210293.

Heyes, C. (2023). Rethinking norm psychology. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, 17456916221112075.

House, B. R., Kanngiesser, P., Barrett, H. C., Broesch, T., Cebioglu, S., Crittenden, A. N., …, Silk, J. B. (2019). Universal norm psychology leads to societal diversity in prosocial behaviour and development. *Nature Human Behaviour*, *4*(1), 36–44.

House, B. R., Silk, J. B., Henrich, J., Barrett, H. C., Scelza, B. A., Boyette, A. H., …, Laurence, S. (2013). Ontogeny of prosocial behavior across diverse societies. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(36), 14586–14591.

House, B. R., Kanngiesser, P., Barrett, H. C., Yilmaz, S., Smith, A. M., Sebastian-Enesco, C., …, Silk, J. B. (2020). Social norms and cultural diversity in the development of third-party punishment. *Proceedings. Biological Sciences/The Royal Society*, *287*(1925), 20192794.

Kallgren, C. A., Reno, R. R., & Cialdini, R. B. (2000). A focus theory of normative conduct: When norms do and do not affect behavior. *Personality & Social Psychology Bulletin*, *26*(8), 1002–1012.

Kandler, A., & Laland, K. N. (2009). An investigation of the relationship between innovation and cultural diversity. *Theoretical Population Biology*, *76*(1), 59–67.

Marlowe, F. W. (2005). Hunter–gatherers and human evolution. *Evolutionary Anthropology*, *14*(2), 54–67.

Mathew, S., Boyd, R., & Van Veelen, M. (2013). Human cooperation among kin and close associates may require enforcement of norms by third parties. In *cultural evolution* (pp. 45–60). The MIT Press.

McElreath, R., Lubell, M., Richerson, P. J., Waring, T. M., Baum, W., Edsten, E., …, Paciotti, B. (2005). Applying evolutionary models to the laboratory study of social learning. *Evolution and Human Behavior: Official Journal of the Human Behavior and Evolution Society*, *26*(6), 483–508.

Molleman, L., Quiñones, A. E., & Weissing, F. J. (2013). Cultural evolution of cooperation: The interplay between forms of social learning and group selection. *Evolution and Human Behavior: Official Journal of the Human Behavior and Evolution Society*, *34*(5), 342–349.

North, D. C. (2010). *Understanding the process of economic change*. Princeton University Press.

O'Gorman, R., Wilson, D. S., & Miller, R. R. (2008). An evolved cognitive bias for social norms. *Evolution and Human Behavior: Official Journal of the Human Behavior and Evolution Society*, *29*(2), 71–78.

Parsons, T. (1937). *The structure of social action*. McGraw-Hill.

Parsons, T. (1951) *The social system*. Free Press.

Peña, J., Volken, H., Pestelacci, E., & Tomassini, M. (2009). Conformity hinders the evolution of cooperation on scale-free networks. *Physical Review E*, *80*(1), 016110.

Perkins, H. W., & Berkowitz, A. D. (1986). Perceiving the community norms of alcohol use among students: some research implications for campus alcohol education programming. *The International Journal of the Addictions*, *21*(9–10), 961–976.

Rakoczy, H., Warneken, F., & Tomasello, M. (2008). The sources of normativity: Young children's awareness of the normative structure of games. *Developmental Psychology*, *44*(3), 875–881.

Rilling, J. K., Sanfey, A. G., Aronson, J. A., Nystrom, L. E., & Cohen, J. D. (2004). The neural correlates of theory of mind within interpersonal interactions. *NeuroImage*, *22*(4), 1694–1703.

Romano, A., & Balliet, D. (2017). Reciprocity outperforms conformity to promote cooperation. *Psychological Science*, *28*(10), 1490–1502.

Sandholm, W. H. (2010). *Population games and evolutionary dynamics*. MIT Press.

Schultz, P. W., Nolan, J. M., Cialdini, R. B., Goldstein, N. J., & Griskevicius, V. (2007). The constructive, destructive, and reconstructive power of social norms. *Psychological Science*, *18*(5), 429–434.

https://doi.org/10.1017/ehs.2024.37 Published online by Cambridge University Press

Smith, D. (2020). Cultural group selection and human cooperation: a conceptual and empirical review. *Evolutionary Human Sciences*, *2*, e2.

Sutter, M., & Kocher, M. G. (2007). Trust and trustworthiness across different age groups. *Games and Economic Behavior*, *59*(2), 364–382.

van de Waal, E., Borgeaud, C., & Whiten, A. (2013). Potent social learning and conformity shape a wild primate's foraging decisions. *Science*, *340*(6131), 483–485.

Whiten, A., Horner, V., & de Waal, F. B. M. (2005). Conformity to cultural norms of tool use in chimpanzees. *Nature*, *437*(7059), 737–740.

Cite this article: Kido Y, Takezawa M (2024). Coevolution of norm psychology and cooperation through exapted conformity. *Evolutionary Human Sciences* **6**, e35, 1–18. https://doi.org/10.1017/ehs.2024.37