Evolution and Human Behavior 42 (2021) 469–479

Contents lists available at ScienceDirect

Evolution and Human Behavior

journal homepage: www.elsevier.com/locate/ens

Do descriptive social norms drive peer punishment? Conditional punishment strategies and their impact on cooperation

Xueheng Li [a,b,1], Lucas Molleman [b,c,d,*,1], Dennie van Dolder [b,e,f,1]

[a] Economics Experimental Lab, Nanjing Audit University, China
[b] Centre for Decision Research and Experimental Economics, University of Nottingham, United Kingdom
[c] Amsterdam Brain and Cognition, University of Amsterdam, the Netherlands
[d] Center for Adaptive Rationality, Max Planck Institute for Human Development Berlin, Germany
[e] School of Business and Economics, Vrije Universiteit Amsterdam, the Netherlands
[f] Tinbergen Institute, the Netherlands

ARTICLE INFO

Keywords:
Cooperation
Peer punishment
Decision-making experiment
Sanctioning
Online experiment
Conditional strategies

ABSTRACT

Peer punishment is widely considered a key mechanism supporting cooperation in human groups. Although much research shows that human behavior is shaped by the prevailing social norms, little is known about how punishment decisions are impacted by the social context. We present a set of large-scale incentivized experiments in which participants (999 American participants recruited via Amazon Mechanical Turk) could punish their partner conditional on either the level of cooperation or the level of punishment displayed by others who previously interacted in the same setting. While many participants punish independently of levels of cooperation or punishment, a substantial portion punishes free riding more severely when cooperation is more common ('norm enforcement'), or when free riding is more severely punished by others ('conformist punishment'). With a dynamic model we demonstrate that conditional punishment strategies can substantially promote cooperation. In particular, conformist punishment helps cooperation to gain a foothold in a population, and norm enforcement helps to maintain cooperation at high levels. Our results provide solid empirical evidence of conditional punishment strategies and illustrate their possible implications for the dynamics of human cooperation.

1. Introduction

For organizations, communities, and society as a whole to function, individuals often have to engage in activities that are costly for themselves, but beneficial for others. Peer punishment is considered to be one of the key mechanisms explaining the emergence and maintenance of cooperation in situations where private and collective incentives do not align (Axelrod, 1986; Boyd, Gintis, & Bowles, 2010; Boyd & Richerson, 1992; Hauert, Traulsen, Brandt, Nowak, & Sigmund, 2007; Sigmund, 2007; Yamagishi, 1986). Empirical evidence shows that many people are willing to punish those who free ride on the cooperation of others, even if punishment is costly and cannot lead to future benefits (Balafoutas, Nikiforakis, & Rockenbach, 2014; Fehr & Gächter, 2000; Fehr & Schurtenberger, 2018; Guala, 2012; Molleman, Kölle, Starmer, & Gächter, 2019; Ostrom, Walker, & Gardner, 1992). The threat of punishment makes free riding less attractive and can thereby help support cooperation at high levels (Arechar, Gaechter, & Molleman, 2018; Crockett, Clark, Lieberman, Tabibnia, & Robbins, 2010; Cubitt, Drouvelis, & Gächter, 2011; Egas & Riedl, 2008; Fehr, Fischbacher, & Gächter, 2002; Fehr & Gächter, 2000; Gächter, Renner, & Sefton, 2008; Nikiforakis, 2010; Nikiforakis & Normann, 2008; Raihani, Thornton, & Bshary, 2012; Rand, Dreber, Ellingsen, Fudenberg, & Nowak, 2009; Rockenbach & Milinski, 2006).

Given that peer punishment can play a pivotal role in sustaining cooperation, it is critical to understand what factors influence people's willingness to punish. When studying the drivers of peer punishment, laboratory studies typically focus on aspects specific to the interaction at hand, such as peers' cooperation decisions, the cost and impact of punishment, or the potential for future interaction or retaliation (Camera & Casari, 2009; Egas & Riedl, 2008; Fehr & Gächter, 2000; Gächter, Kölle, & Quercia, 2017; Nikiforakis, 2010; Raihani & Bshary, 2019). In doing so, these studies generally abstract away from the broader social

* Corresponding author at: Nieuwe Achtergracht 129-B, 1018 WT Amsterdam, the Netherlands.
E-mail address: l.s.molleman@uva.nl (L. Molleman).
[1] Equal contributions

https://doi.org/10.1016/j.evolhumbehav.2021.04.002
Received 17 October 2020; Received in revised form 30 March 2021; Accepted 11 April 2021
Available online 30 April 2021

context in which an interaction takes place. Cross-cultural experiments, however, show that social context matters for the effectiveness of punishment to support cooperation: people from different societies use peer punishment in systematically different ways (Gächter & Herrmann, 2009; Henrich, 2000; Henrich, 2016; Henrich et al., 2010, 2006; Herrmann, Thöni, & Gächter, 2008; Oosterbeek, Sloof, & van de Kuilen, 2004; Roth, Prasnikar, Okuno-Fujiwara, & Zamir, 1991). Because societies differ from each other in myriad ways, such cross-cultural comparisons have limited ability to identify exactly which aspects of the social context underlie any observed differences.

In this paper, we investigate an important way in which the social context may influence punishment of free riding: by indicating 'descriptive norms' specifying what behavior is typical in the current interaction setting (Berkowitz, 2005; Bicchieri, 2006; Cialdini, Kallgren, & Reno, 1991). Studies from across the social sciences have shown that people tend to conform to descriptive norms (Bond & Smith, 1996; Burger et al., 2010; Cialdini et al., 1991; Cialdini, Reno, & Kallgren, 1990; Frey & Meier, 2004; Hallsworth, List, Metcalfe, & Vlaev, 2017; Nolan, Schultz, Cialdini, Goldstein, & Griskevicius, 2008). In social dilemmas, it has been established that many people are more willing to cooperate if they believe that others will do so as well (Bicchieri, 2006; Elster, 1989a, 1989b; Fehr & Schurtenberger, 2018; Fischbacher, Gächter, & Fehr, 2001; Gächter et al., 2017; Henrich, 2016). Whether, and if so how, descriptive norms influence peer punishment, however, remains unclear. Here, we first conduct incentivized experiments and show that many people condition their punishment of a free-riding partner on descriptive norms of cooperation and punishment. With a simple dynamic model, we then show that such conditional punishment strategies can have pronounced implications for the emergence and maintenance of cooperation in groups.

For the decision to punish a free riding peer, two descriptive norms may be important. First, punishment decisions might be guided by the descriptive norm of cooperation: is free riding the typical action in the population? It has been shown that people often infer injunctive norms (what one ought to do) from descriptive norms (what most people actually do). People tend to judge behaviors that are less common in a population to be less socially appropriate (or 'moral') and consequently more deserving of punishment (Chudek & Henrich, 2010; Eriksson, Cownden, Ehn, & Strimling, 2014; FeldmanHall, Otto, & Phelps, 2018; Hume, 2003; Kelley, 1971; Lindström, Jangard, Selbing, & Olsson, 2018; McGraw, 1985; Son, Bhandari, & FeldmanHall, 2019; Tworek & Cimpian, 2016; Welch et al., 2005). If people use descriptive norms of cooperation to form moral judgments in this manner, they will judge free riding more harshly when it is atypical, which will increase their willingness to punish. Second, punishment decisions might be guided by the descriptive norm of punishment: is punishment a typical reaction to free riding? Descriptive norms of punishment can signal a 'principle of social proof' (Cialdini & Trost, 1998) that free riding is disapproved of, and that punishment is an appropriate and legitimate reaction. Conformity to these norms would lead people to punish free riding if others do so as well. Examining the impact of these two descriptive norms on sanctioning behavior increases our understanding of how the social context can affect individuals' punishment of free riding and thereby influence the emergence and maintenance of cooperation.

To investigate whether descriptive norms of cooperation and punishment impact peer punishment, we conduct two decision-making experiments. Participants are randomly paired and play a prisoner's dilemma with punishment. Our implementation consists of two stages. In the first stage, participants decide to either 'cooperate' or 'defect'. In the second stage, they decide how severely they want to punish their partner if their partner chose to defect. We add minimal social context by allowing participants to condition their punishment decisions on the levels of cooperation and punishment displayed by a random sample of participants who previously interacted in the same setting (hereafter, the 'reference group'). In one experiment, participants can condition their punishment decisions on the level of cooperation in the reference group; in the other experiment, participants can condition their punishment on the level of punishment in the reference group. Importantly, the decisions of members of the reference group do not affect payoffs of the focal participants.

Our setup enables us to classify individual participants according to how their punishment decisions respond to descriptive social norms, thereby deepening empirical understanding of individual differences in (conditional) punishment. Individual differences in conditional cooperation have received considerable attention in prior research, indicating that the dynamics of cooperation in groups strongly depend on the interplay of individuals' conditional strategies and their beliefs about others' cooperativeness (Fischbacher et al., 2001; Fischbacher & Gachter, 2010; Weber, Weisel, & Gächter, 2018). In contrast, little is known about individual differences in conditional punishment and the way in which these differences may affect the emergence of cooperation. Our experimental design allows us to isolate the possible effects of descriptive norms on punishment from related considerations such as a preference for coordinated punishment or positive reciprocity towards other punishers (Casari & Luini, 2009, 2012; Guala, 2012; Kamei, 2014; Molleman et al., 2019). Finally, by creating controlled conditions that systematically differ in terms of descriptive norms of cooperation and punishment, our setup complements cross-cultural experiments on punishment that rely on natural variation in social context (Gächter & Herrmann, 2009; Henrich, 2016; Henrich et al., 2001, 2010, 2006; Herrmann et al., 2008; Oosterbeek et al., 2004; Roth et al., 1991).

Our results demonstrate that on aggregate, people's willingness to punish their free riding partner increases both with the level of cooperation and with the level of punishment in the reference group. Importantly, we observe substantial heterogeneity in how people react to the level of cooperation and the level of punishment. When participants can condition their punishment on the fraction of cooperators in the reference group, many engage in 'norm enforcement', punishing their partner more with increasing cooperation levels. When participants can condition their punishment on the level of punishment in the reference group, many engage in 'conformist punishment', punishing their partner more with increasing punishment levels. In both cases, a substantial fraction of participants engages in 'independent punishment', applying the same punishment intensity irrespective of the descriptive norm.

To examine the possible long-term implications of the experimentally observed conditional punishment strategies, we develop a simple dynamic model in which a population of agents recurrently interact in a social dilemma game with punishment similar to our experiments. We use analytical methods and agent-based simulations to evaluate how the experimentally observed punishment strategies can shape cooperation in a population.

The model captures key qualitative features of social norm dynamics, involving prolonged periods of stability and sudden shifts (Young, 2015). Moreover, the model shows that, in conjunction with independent punishers, norm enforcement and conformist punishment can effectively support cooperation. Importantly, we find that norm enforcement and conformist punishment play markedly different roles in promoting cooperation: conformist punishment can effectively promote the establishment of cooperation in a population, whereas norm enforcement is particularly effective at maintaining cooperation at high levels. Overall, our model shows that the experimentally observed conditional punishment strategies can have a strong and positive impact on the dynamics of cooperation.

2. Material and methods

2.1. Experimental design

We randomly matched participants in pairs to play a two-stage game in which they could earn points. Participants received an initial endowment of 25 points. In the first stage, the two players

simultaneously choose to cooperate or defect. Joint payoffs are highest when both partners cooperate, with both earning 18 points. However, each individual can increase their personal payoffs in this stage by choosing to defect: unilateral defection leads to 25 points for self and 9 points for the other. Mutual defection leads to 16 points for each. In the second stage, participants have the opportunity to punish their interaction partner if their partner chose to defect (by design excluding 'antisocial punishment'; see Herrmann et al. 2008), by assigning up to 10 deduction points to them. Each assigned deduction point reduces the participant's payoffs with 1 point, and the partner's payoffs with 3 points. The participants' total payoff thus constituted the endowment of 25 points plus the points earned in Stage 1 minus the potential costs of conducting punishment and the potential losses from being punished in Stage 2. The Nash equilibrium of this one-shot game is to defect in the first stage, and to not assign any deduction points in the second stage.

We report on two experiments (total $N = 999$) in which participants could condition their punishment on descriptive norms of cooperation (CC experiment; $N = 498$) or descriptive norms of punishment (CP experiment; $N = 501$). We operationalized these descriptive norms as behavior in a reference group of individuals who previously interacted in the same setting, but who were irrelevant for the payoffs in the current interaction. Participants had to indicate how many deduction points they would assign to their partner (if the partner chose to defect) for a set of situations that vary with respect to the reference group's levels of cooperation or punishment. The actual behavior in the reference group determined which of the situations was implemented and used to calculate payoffs. Section 2.3. describes our measurement of conditional punishment.

2.2. Experimental procedures

We recruited participants by posting Human Intelligence Tasks ('HITs') on Amazon Mechanical Turk (MTurk) during September 2017 and September 2019. We restricted our sample to the United States for reasons of comprehension of English instructions. The only other participation criterion was to have at least 95% of previous HITs approved. The average age in our sample was 35.5 (s.d. = 10.3, range 18–71) and 43% of participants was male.

The experiments were programmed in LIONESS Lab (Giamattei, Yahosseini, Gächter, & Molleman, 2020), code is available via the public repository associated with this paper (https://github.com/LucasMollemman/LMD_Conditional_punishment); experimental instructions are documented in full in the Appendix, Section 4. Ethical approval was given by the Research Ethics Committee at the School of Economics, University of Nottingham, UK.

After reading the instructions and passing compulsory control questions, participants entered Stage 1 and made their binary cooperation decisions. In Stage 2, participants completed another set of compulsory control questions, before we asked them to provide their punishment responses to descriptive norms of cooperation and punishment (see Section 2.3). Once participants had completed the two decision making stages of the experiment, they were placed in a waiting room, in which they would be matched with another participant as soon as they completed their decisions as well. In case no match could be made within 5 min, participants could choose to leave and receive a fixed bonus payment of $1.00, or to wait for another 2 min for a possible matching partner (as in Arechar et al., 2018; see Table A1 for details on dropouts during the experiment). Participants were informed that from the point of reaching the waiting room onwards, they did not have to make any further decisions. Excluding the time spent in the waiting room, our experiments on average lasted 9.9 min. The points earned by participants were converted to US dollars at the end of the experiment (20 points were worth $1.00). Average earnings were $1.96 (range $0.41 - $2.51), which translates to an hourly wage of approximately $12.00.

2.3. Measurement of conditional punishment

In the CC experiment, we operationalized the descriptive norm of cooperation as the fraction of cooperative choices in a payoff-irrelevant reference group (sampled from a pre-recorded pool; see below). We used the 'strategy method' (Selten, 1967) and presented participants with eleven situations regarding the proportion of cooperators in this reference group, spanning the full range of possible outcomes. For each of these situations, participants had to indicate how many deduction points they would assign to their current interaction partner, if their partner chose to defect. In the CP experiment, we operationalized the descriptive norm of punishment as the average intensity of punishment in the reference group, and participants indicated for each possible situation how many deduction points they would assign to their current interaction partner.

The pre-recorded pool consisted of a total of 273 MTurkers who played a prisoner's dilemma with punishment mirroring our experiments (cooperation rate: 69%; average punishment of free riding partners: 2.7 deduction points). For each dyad in the main experiments, we independently sampled 50 participants from the pre-recorded pool to form the reference group. The behavior of the reference group defined the situation that was used to calculate participants' earnings. Since participants did not know which situation was the actual one beforehand, they had an incentive to consider each situation as if it was real.

Using the strategy method has the advantage that it yields a full profile of punishment responses for each participant. This approach is commonly used to study how cooperation depends on the cooperativeness of interaction partners (Fischbacher et al., 2001; Herrmann & Thöni, 2009), and has proven useful in studying the determinants of punishment in cooperative interactions (Brandts & Charness, 2011; Cheung, 2014; Falk, Fehr, & Fischbacher, 2005; Kamei, 2014; Molleman et al., 2019). Although there is some concern that the strategy method may amplify the degree of conditionality observed (Burton-Chellew, El Mouden, & West, 2016; Ferraro et al., 2010; Columbus & Böhm, 2021), empirical research on cooperation has shown that participants who condition their behavior on the decisions of others under the strategy method also do so under the direct response method (Fischbacher, Gächter, & Quercia, 2012).

3. Results

3.1. Experimental results

The experiments did not differ in terms of cooperation rates (CC: 68.5%; CP: 65.5%; $X^2(1) = 1.02$, $P = 0.313$). Participants' overall punishment levels, averaged across all situations, were also similar across experiments (CC: 2.68; CP: 2.47 deduction points; two-sample Wilcoxon rank-sum test: df = 997, $z = 1.703$, $P = 0.089$).

On aggregate, behavior in the reference group impacted the participants' punishment decisions (Fig. 1). The fraction of cooperators had a small but significant positive effect on the average number of deduction points that participants assigned to their free riding partners (Fig. 1a; $P = 0.002$; see Table A2, Model 1 for ordinary least squares regression results). The average intensity of punishment had a stronger impact on average punishment (Fig. 1b; $P < 0.001$; Table A2, Model 2). We interpret this as evidence that the social context impacts peer punishment, with both descriptive norms of cooperation and descriptive norms of punishment modulating people's overall willingness to punish defectors.

Participants substantially differed in their punishment behavior (Fig. 2). In the CC experiment, 64% of participants punished at least once. Thirty-six percent of participants never punished, thereby avoiding the costs of punishment and maximizing their own payoff. Among the punishers, we observe that two distinct strategies predominate (Fig. 2a). A large portion of punishers (33.7%) engages in 'norm enforcement', monotonically increasing punishment with the level of

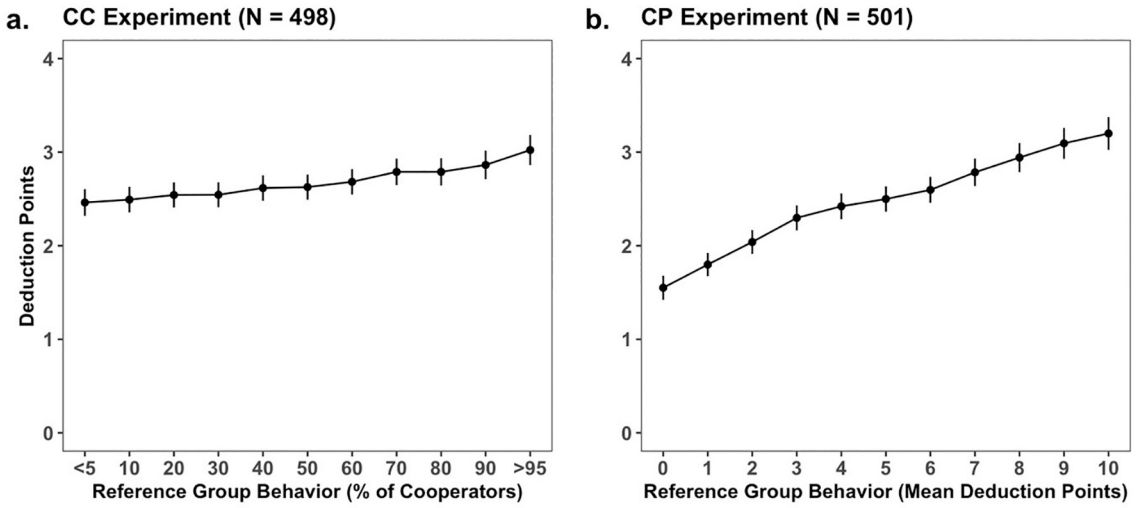

Fig. 1. Average intensity of punishment as a function of cooperation and punishment among participants in the payoff-irrelevant reference group. Panels **a** and **b** summarize decisions in the CC and CP experiment, respectively, showing the average deduction points for each of the situations presented to participants (see screenshots of punishment stage in the experiment). Error bars indicate standard errors of the means (SEM). For statistical analysis, see Table A2.

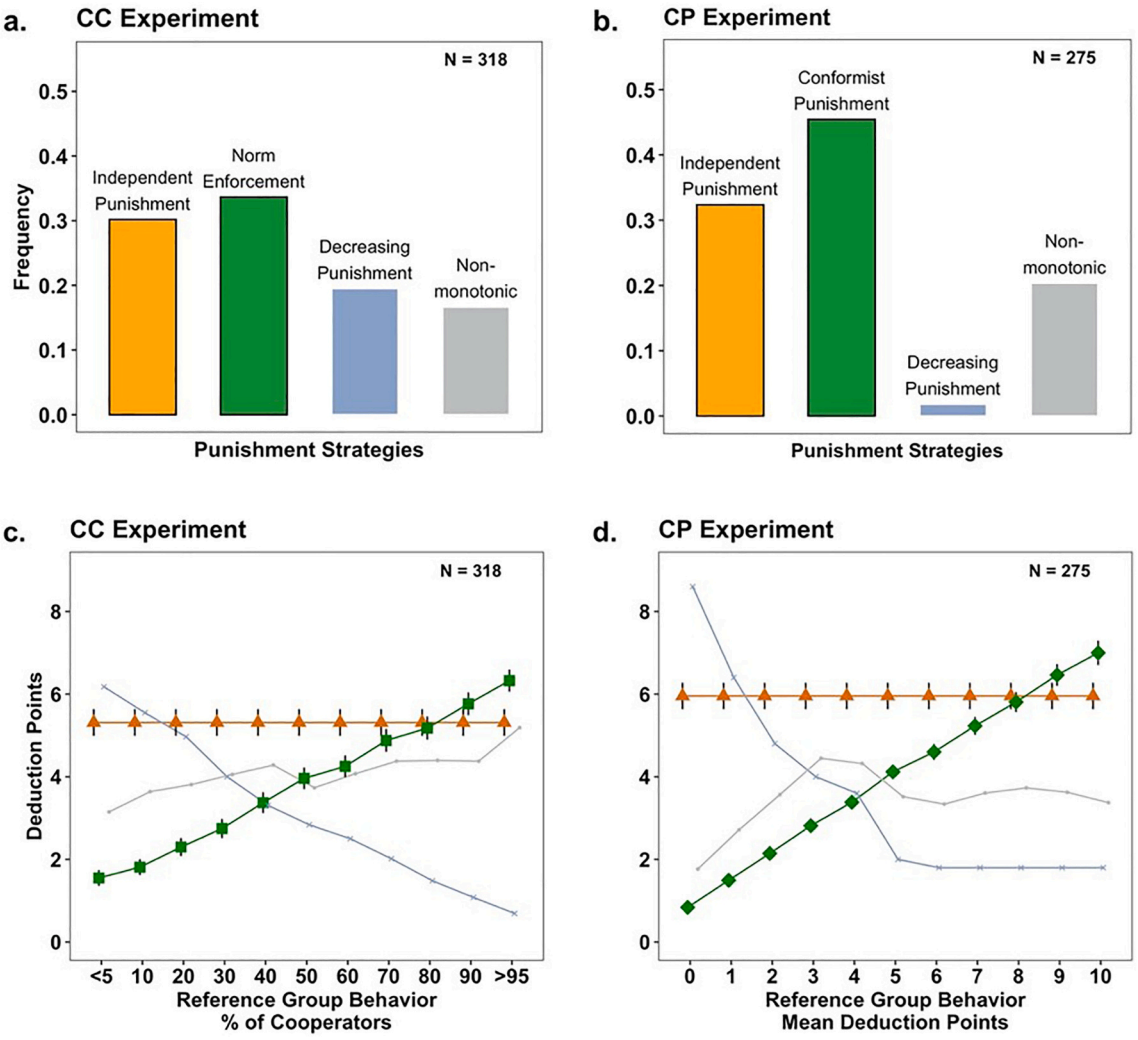

Fig. 2. Punishment strategies observed in our experiment. **a** and **b**, Frequency distributions of punishment strategies in the CC and CP experiment, among participants who punish at least once (64% and 54% in the CC and CP experiment, respectively). **c** and **d**, For each strategy, the average number of deduction points (+/− 1 SEM) assigned to free riding partners for each of the situations in the CC and CP experiment.

cooperation in the reference group (Fig. 2a; green bar). An approximately equal portion (30.2%) applies 'independent punishment', responding with the same punishment intensity irrespective of the level of cooperation in the reference group (Fig. 2a; orange bar). A smaller portion of punishers (19.5%) decreased their punishment of free riders as cooperation became more common in the reference group (Fig. 2a; blue bar).

In the CP experiment, 55% of participants punished at least once (and 45% never punished, maximizing their own payoffs). Again, we observe that among these punishers, two distinct strategies predominate (Fig. 2b). A large portion of punishers (45.5%) engaged in 'conformist punishment', monotonically increasing punishment with the level of punishment in the reference group (Fig. 2b; green bar). A substantial portion (32.4%) applied 'independent punishment', responding with the same punishment intensity irrespective of punishment in the reference group (Fig. 2a; orange bar). 'Decreasing punishment' was virtually absent in the CP experiment (Fig. 2b, blue bar).

In both experiments, the distribution of strategies did not depend on the number of attempts participants required to complete the control questions (Fig. A1 and A2), suggesting that the observed patterns did not stem from participants misunderstanding the instructions of the decision situation. Taken together, these results indicate that people substantially vary in how they condition punishment of free riders on the levels of cooperation and punishment in the social environment.

Fig. 2c and d show, for each of the punishment strategies, the average number of deduction points assigned to free riding partners for each of the situations in the CC experiment (Fig. 2c) and the CP experiment (Fig. 2d). Although behavior in reference group had only modest effects on the aggregate level of punishment, the subset of participants who engaged in norm enforcement (Fig. 2c; green) and conformist

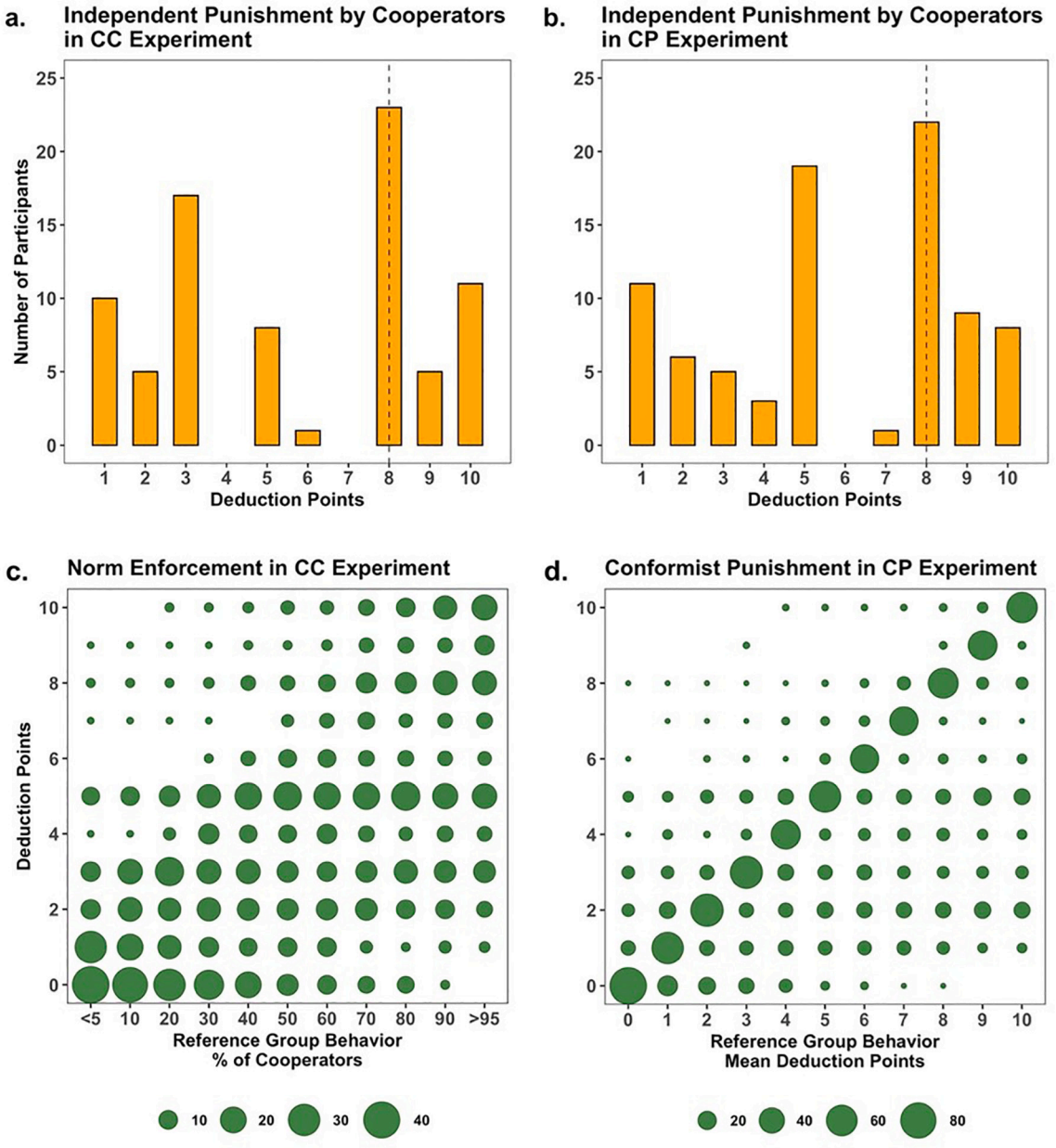

Fig. 3. Punishment behavior for the most common punishment strategies. **a** and **b,** Distributions of deduction points assigned by participants who punished independently and cooperated in the first stage of the game. The mode behavior for both experiments (assigning 8 deduction points; vertical dotted line) equalizes the earnings of a cooperator and a free rider. **c** and **d,** Deduction points assigned by participants who engaged in 'norm enforcement' in the CC experiment, and 'conformist punishment' in the CP experiment. Dot sizes reflect the numbers of observations.

punishment (Fig. 2d; green) strongly reacted to the level of cooperation and punishment in the reference group. On average, 'norm enforcing' participants assigned 1.6 deduction points when the percentage of cooperators in the reference group was less than 5%. Their punishment increased to 6.3 deduction points when more than 95% of the participants in the (payoff-irrelevant) reference group cooperated (Fig. 2c; green line). Similarly, in the CP experiment, participants who punished conformistically assigned about 0.8 deduction points when participants in the reference group assigned 0 deduction points on average. Their punishment increased to 6.5 deduction points when the average number of deduction points assigned by members of the reference group was 10. These results show that the punishment behavior of participants who use conditional strategies is strongly affected by the social environment.

For participants who punished independently and cooperated in Stage 1, the modal behavior in both experiments was to assign 8 deduction points (Fig. 3a,b). By contrast, assigning 8 deduction points is very rare among independent punishers who defected in Stage 1 (see Fig. A3–6 for a full breakdown of punishment decisions by cooperators and defectors in each experiment). This level of punishment equalizes the earnings between a cooperator and their free-riding partner, suggesting that some participants' do not punish to reciprocate the unkind action, but rather to eliminate disadvantageous inequality (Fehr & Schmidt, 1999; Raihani & Bshary, 2019).

Fig. 3c and d show the distributions of assigned deduction points among participants engaging in norm enforcement in the CC experiment (Fig. 3c) and conformist punishment in the CP experiment (Fig. 3d). We observe large numbers of data points on the diagonal in the graph for conformist punishment in the CP experiment. This indicates that participants engaging in conformist punishment frequently chose to exactly match the average number of deduction points assigned in the reference group. Norm enforcement in the CC experiment showed a less pronounced pattern.

3.2. Dynamic model

Our experimental results reveal that people's punishment of free riders is shaped by descriptive norms of cooperation and punishment, and that various punishment strategies (conditional and unconditional) co-exist. This raises the question of how the observed punishment strategies interact to drive dynamics of cooperation over time. To address this question, we develop a simple model and evaluate it using analytical methods and agent-based simulations. In line with the existing theoretical literature on dynamics of cooperation and punishment (Boyd, Gintis, Bowles, & Richerson, 2003; Henrich & Boyd, 1998; Szolnoki & Perc, 2013), we abstract away from many details of real-world interactions to focus on fundamental mechanisms and their basic implications for cooperation. In particular, we consider simplified versions of the conditional and unconditional punishment strategies observed in our experiment, and examine how their relative frequencies in a population impact the emergence and maintenance of cooperation.

We consider a fixed population of n agents who interact repeatedly for T periods in a setting similar to our experiment. In each period, agents (*i*) are randomly matched into pairs, (*ii*) choose whether to cooperate or defect, and (*iii*) choose whether to punish their partner if their partner defects. For ease of exposition and to facilitate tractability, we model both cooperation and punishment as binary decisions (see Appendix, Section 3 for discussion). In each period, each agent samples m agents from the population and counts how many of them cooperated and how many of them were willing to punish defectors in the previous period. The counts divided by m reflect their beliefs about the rates of cooperation and punishment in the current period (respectively denoted by b_c and b_p). An agent cooperates in the current period if they believe that the proportion of punishers exceeds a threshold ($b_p > \theta_C$), and defects otherwise.

An agent's punishment strategy determines whether they punish a defecting partner. Based on our experimental results, we consider four punishment strategies: (*i*) *independent punishment*: punish irrespective of beliefs; (*ii*) *norm enforcement*: punish if the perceived cooperation rate exceeds a threshold ($b_c > \theta_{NE}$); (*iii*) *conformist punishment*: punish if the perceived punishment rate exceeds a threshold ($b_p > \theta_{CP}$); and (*iv*) *never punish*. For the sake of exposition, we will here focus on the case where $\theta_C = \theta_{NE} = \theta_{CP} = 0.5$. In the Supplementary Analysis (Propositions 1 and 2 in the Appendix), we consider arbitrary threshold values, which lead to qualitatively similar results. To compare the effects of punishment strategies on cooperation, we assume that agents' punishment strategies are fixed throughout all periods of the simulation, and are mutually exclusive (that is, each agent is characterized by a single punishment strategy).

The dynamics are stochastic: with probability $\varepsilon > 0$, an agent makes a mistake and behaves randomly; with complementary probability $1 - \varepsilon$ the agent behaves according to its strategy (Kandori, Mailath, & Rob, 1993; Young, 1993). Mistakes are independent across agents, periods, and cooperation and punishment decisions.

Our goal is to assess how relative frequencies of independent punishment (Q_{IP}), norm enforcement (Q_{NE}), and conformist punishment (Q_{CP}) affect the dynamics of cooperation. First, we derive analytical results about the stationary distribution of the dynamic when the observation sample is large ($m = n$) and the mistake probability is vanishingly small ($\epsilon \to 0$). The stationary distribution reflects the relative frequencies of different population states in the long run ($T \to \infty$). We show that if $Q_{IP} + \frac{1}{2}(Q_{NE} + Q_{CP}) > \frac{1}{2}$, only the cooperation equilibrium occurs with a positive frequency in the stationary distribution. Conversely, if $Q_{IP} + \frac{1}{2}(Q_{NE} + Q_{CP}) < \frac{1}{2}$, only the defection equilibrium occurs with a positive probability in the stationary distribution (see Appendix, Section 3 for formal proof). This analysis shows that, perhaps unsurprisingly, independent punishment is the most potent strategy for promoting cooperation. Importantly, however, when independent punishment is not sufficiently frequent, conditional strategies of norm enforcement and conditional punishment can be key for sustaining cooperation in the long run.

Next, we use simulations to examine the short-run dynamics of our model: how conditional punishment strategies drive the emergence and breakdown of cooperation, and how their relative frequencies affect the time it takes for a population to transition between states of high and low cooperation. Simulations also allow us to consider small observation samples and non-negligible mistake probabilities. To account for path-dependence, the simulations consider different starting conditions by varying agents' initial beliefs about the rates of cooperation and punishment. To evaluate how norm enforcement and conformist punishment affect cooperation, we fix the frequency of independent punishers at 30% and vary the frequencies of the conditional punishment strategies. Further robustness checks are detailed at the end of this section. The full simulation code is available from the public repository associated with this paper (https://github.com/LucasMolleman/LMD_Conditional_punishment).

Fig. 4 shows the dynamics of cooperation in situations where independent punishment is not sufficiently frequent to sustain cooperation by itself. We first confirm that, if independent punishers alone are too rare to support cooperation on their own, and neither of the conditional punishment strategies is present in the population, cooperation never emerges in our simulations (Fig. 4a,b). Next, we consider cases where independent punishment is complemented with conditional punishment strategies, raising the overall frequency of punishers. The presence of norm enforcement has a strong stabilizing effect once high levels of cooperation have been achieved (Fig. 4c). However, it might take considerable time for cooperation to emerge (Fig. 4d). These dynamics are driven by a positive feedback loop between norm enforcement and cooperation, locking a population into a state of either high or low cooperation, making it hard to transition from one state to the other.

By contrast, in the presence of conformist punishers cooperation readily emerges, but is not stable (Fig. 4e,f). The population alternates

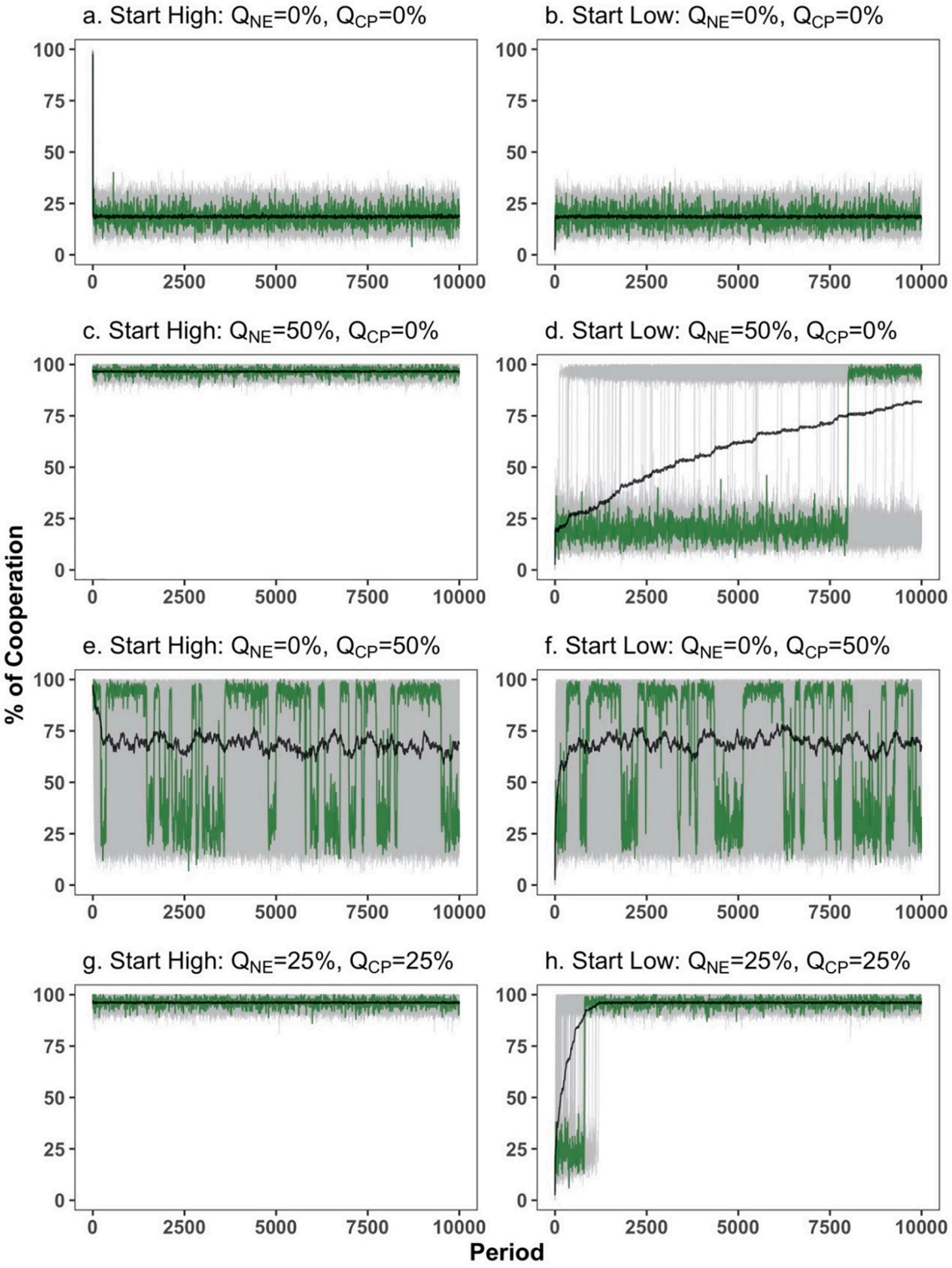

Fig. 4. Effects of conditional punishment strategies on cooperation dynamics. Across all panels, we hold fixed the frequency of independent punishers at 30%. Q_{NE} is the frequency of norm enforcement, and Q_{CP} is the frequency of conformist punishment. Columns of panels vary agents' initial beliefs regarding the frequencies of punishment and cooperation in the population. In the left-hand side column, these beliefs start high ($b_c = b_p = 0.75$). This means that for all agents, the payoff maximizing response in the first period is to cooperate; independent punishers, norm enforcers, and conformist punishers punish their defecting partner when holding these beliefs. In the right-hand side column, beliefs start low ($b_c = b_p = 0.25$). This means that for all agents, the payoff maximizing response in the first period is to defect; only independent punishers punish defectors when holding these beliefs. In each period after the first, each agent updates their beliefs with probability u. We set $u = 0.5$ in our simulations; our analytical results apply to any $0 < u < 1$ (see Remarks in the Appendix, Section 3.4). In each panel, black lines show mean cooperation rates over time across 100 simulation runs; grey lines show individual runs, with a representative run highlighted in green. Further simulation settings: $n = 100$, $m = 10$, $\varepsilon = 0.05$. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

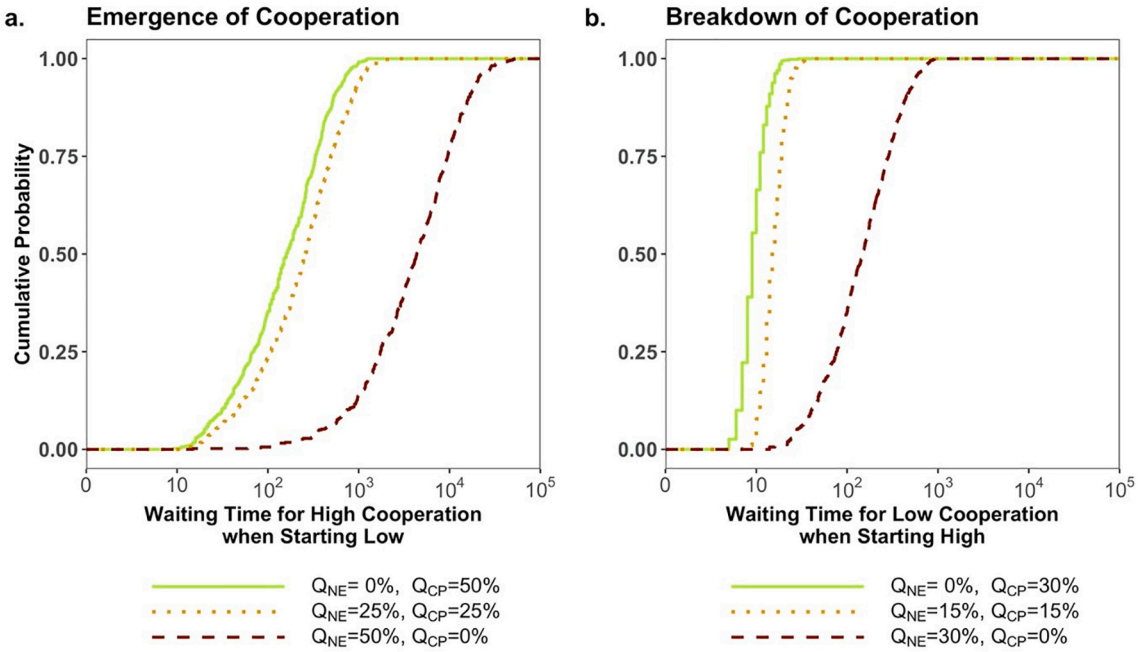

Fig. 5. Effects of conditional punishment strategies on the emergence and breakdown of cooperation. Lines show the cumulative probability of cooperation to rise above 75% (**a**) or fall below 25% (**b**), as a function of time. Time is shown on a logarithmic scale, and each line represents 500 simulation runs. Across both panels, we hold fixed the frequency of independent punishers at 30%. Q_{NE} is the frequency of norm enforcement, and Q_{CP} is the frequency of conformist punishment. Frequencies of these conditional strategies were chosen such that—according to our analytical results—cooperation would emerge (Panel **a**) or break down (panel **b**) in the long run. Initial beliefs regarding cooperation and punishment levels start low in Panel **a** ($b_c = b_p = 0.25$), and high in Panel **b** ($b_c = b_p = 0.75$; see Methods for details). Each simulation runs for 100,000 (10^5) periods. Further simulation settings: $n = 100$, $m = 10$, $\varepsilon = 0.05$. Results for additional population compositions with regard to punishment strategies confirm the general pattern shown here (Fig. A7).

between states with low and high levels of cooperation, with rapid shifts between these states. These dynamics are driven by another positive feedback loop: when levels of cooperation and punishment are low, some agents may punish their free riding partner due to mistakes or—in the case of conformist punishers—due to sampling bias. In turn, these stochastic events may prompt other conformist punishers to punish too in the next period, thereby increasing the levels of cooperation and punishment even more, and possibly tipping the population to high levels of cooperation and punishment. However, similar stochastic processes may also cause cooperation to suddenly break down when conformist punishers stop punishing because they happen to underestimate the level of punishment in the population.

When both conformist punisherst and norm enforcers are present in the population—but keeping the overall frequency of conditional punishers the same—cooperation rapidly emerges and remains stable at high levels (Fig. 4g,h). Conformist punishers still amplify the impact of stochasticity when cooperation is low, facilitating the emergence of cooperation. Subsequently, norm enforcement locks the population into a state of high cooperation. This result highlights that the concerted action of conformist punishment and norm enforcement can efficiently support cooperation.

These results indicate that different conditional punishment strategies can promote cooperation in different ways: conformist punishment facilitates the emergence of cooperation; norm enforcement helps to maintain it after its emergence. Fig. 5 confirms these insights. When a population starts from a state of low cooperation, the presence of conformist punishment, rather than norm enforcement, can strongly increase the rate at which it shifts to a state of high cooperation (Fig. 5a). Conversely, the presence of norm enforcement can substantially extend the time that a population remains in a state of high cooperation (Fig. 5b).

In the Appendix we examine the generalizability and robustness of our model results. We confirm that our main model results hold across different ranges of relative frequencies of the various (conditional) punishment strategies and different initial beliefs about cooperation and punishment in the population (Fig. A8 and A9). Furthermore, we show that the presence of agents who decrease their punishment of free riding as cooperation becomes more common—as observed in the CC experiment ('decreasing punishment' in Fig. 2a)—destabilizes the non-cooperative equilibrium. By itself, decreasing punishment cannot support high levels of cooperation. However, in conjunction with other conditional punishment strategies, norm enforcement in particular, decreasing punishment can boost the likelihood that a population reaches high and stable levels of cooperation (Fig. A10).

4. Discussion

Our experiments provide behavioral evidence that punishment of free riding in social dilemmas is shaped both by descriptive norms of cooperation ("is free riding a typical action in the population?") and by descriptive norms of punishment ("what is the typical punishment reaction to free riding?"). On aggregate, punishment increases both with the level of cooperation and the level of punishment in a payoff-irrelevant reference group. At the individual level, we observe substantial heterogeneity in how people react to these descriptive norms. Whereas a sizable fraction of participants punishes independently of what others are doing ('independent punishment'), at least as many participants display conditional punishment strategies, increasing their punishment either with higher levels of cooperation ('norm enforcement) or with higher level of punishment ('conformist punishment') in the reference group. Overall, our experimental results support the emerging view that conditional strategies are not limited to the domain of positive reciprocity, i.e., cooperation (Fischbacher et al., 2001; Fischbacher & Gachter, 2010; Keser & Van Winden, 2000; Weber et al., 2018), but are also important in the domain of negative reciprocity, i.e., punishment (FeldmanHall et al., 2018; Kamei, 2014; Molleman et al., 2019; Son et al., 2019).

Our finding that people punish free riding more when cooperation is

more common provides novel behavioral evidence for the idea that people infer injunctive norms (what is 'moral') from descriptive norms (what is 'common'; Chudek & Henrich, 2010; Eriksson et al., 2014; Hume, 2003; Kelley, 1971; Lindström et al., 2018; McGraw, 1985; Tworek & Cimpian, 2016). In doing so, we complement existing research that largely relied on (non-incentivized) moral judgments (Eriksson et al., 2014; McGraw, 1985; Tworek & Cimpian, 2016; Welch et al., 2005). Our finding that people punish free riding more when others do so as well, adds to a growing literature showing that people condition their own punishment decisions on the punishment behavior of others. In previous studies such conformist punishment could often have been driven by preferences for coordinated punishment or positive reciprocity towards other punishers (e.g., Kamei, 2014; Molleman et al., 2019; Kamei, 2020; for a notable exception see Son et al., 2019). Our experimental design rules out such considerations, providing clear evidence for the impact of descriptive social norms on punishment behavior.

Our behavioral approach, however, does not allow us to pin down the psychological mechanisms underlying the different punishment strategies. Recent evidence suggests that norm enforcement may be driven by increased disapproval of free riding when cooperation is common (Lindström et al., 2018). Similarly, conformist punishers' observing others punishing defectors may increase their own disapproval of defection. Alternatively, it could be also that conformist punishers follow a simple heuristic of copying what others are doing (Gigerenzer, 2008, 2010; Lindström et al., 2018). The finding that conformist punishers frequently chose to exactly match the average punishment of others (Fig. 2d) suggests that the latter may be more likely. Future work should combine behavioral data with survey data to investigate to what extent (conditional) punishment reflects changes in people's moral judgments after observing others' actions, and to what extent it reflects people's conformist inclinations.

An important open issue for understanding conditional punishment is whether people who condition their punishment behavior on that of others do so consistently across different decision settings. Although it seems plausible that some individuals are generally more responsive than others to their social environment, it remains an open question whether individuals who engage in norm enforcement when informed of cooperation rates in their environment, would also tend to punish conformistically when informed of punishment rates. Similarly, conditional punishment strategies might correlate with well-studied strategies of conditional cooperation. Experiments addressing these associations would provide deeper insights into the behavioral architecture of cooperation and punishment, contributing to ongoing debates around the generality of strategies across settings involving positive and negative reciprocity (Albrecht, Kube, & Traxler, 2018; Molleman et al., 2019; Peysakhovich, Nowak, & Rand, 2014; Weber et al., 2018).

Our model demonstrates how the experimentally identified conditional punishment strategies can have important implications for cooperation dynamics. Analytical results reveal that conformist punishment strategies can considerably broaden the set of conditions under which cooperation can emerge and persist in the long run. Agent-based simulations yield deeper insights into the different roles that norm enforcement and conformist punishment play in this dynamic. Norm enforcers punish free riders when cooperation rates in the population are relatively high, which makes them effective in maintaining cooperation. However, they do not punish when free riding predominates and are therefore of little help for cooperation to emerge from scratch. In contrast, conformist punishers sanction free riders as long as sufficiently many others do—irrespective of the cooperation rate—and can, therefore, play a valuable role in helping cooperation gain a foothold in a population.

Whereas our study shows that the behavior of an individual can be influenced by what the collective is doing, our model illustrates how these individual strategies can subsequently impact collective dynamics. Previous theoretical studies that incorporated conditional punishment strategies focused on long-run evolutionary dynamics and investigated under which conditions particular punishment strategies are likely to evolve (e.g., Boyd et al., 2010; Szolnoki & Perc, 2013). In contrast, we consider more limited time scales—over which we can take the distribution of strategies in the population as given—and investigate how the relative frequency of particular punishment strategies impact the dynamics of cooperation.

We deliberately employ a simple stylized model to illustrate the basic effects of conditional punishment strategies. Despite its simplifying assumptions (e.g., mutually exclusive punishment strategies, binary punishment and cooperation choices, random re-matching after every interaction), our model produces intuitive and robust results. Moreover, the model is able to capture key qualitative features of the dynamics of social norms: prolonged periods of stability which are punctuated by tipping points, where one norm is rapidly replaced by another (Young, 2015; see Fig. 4). In line with recent results (Lindström et al., 2018), we find that especially the positive social feedback provided by norm enforcers may play an important role in these patterns in norm dynamics.

Our simulations illustrate how conformist punishment can amplify stochastic events, leading to both rapid alternation between the emergence and breakdown of cooperation in a population (Figs. 4, 5). In contrast, norm enforcement can engender a process of positive feedback with cooperation, locking a population into a state of either high or low levels of cooperation, making it hard to transition to the other state (Fig. 4). These results give pointers for efficiently promoting desirable behaviors, such as voting, tax compliance, or energy conservation. In particular, facilitating the observability of (or accessibility to) information about other people's behavior may be effective when the majority of the population displays the desired behavior: this information can boost norm enforcement, ensuring that adherence to the present norm remains high. Conversely, when a majority of the population shows the undesired behavior, it may be more effective to provide people with information that informs them that many people disapprove of the undesirable behavior. Such information may trigger conformist punishment and shift the system to the more desirable outcome.

Declaration of Competing Interest

We declare no conflict of interest.

Acknowledgments

We thank Simon Gächter, Björn Lindström, Jesse Niebaum, Wouter van den Bos and Manwei Liu for useful comments and discussions. We thank Benjamin Beranek for help with data collection. L.M. acknowledges support from the European Research Council (Grant: ERC-AdG 295707 COOPERATION), an Amsterdam Brain and Cognition Project grant 2018, and the Jacobs Foundation. D.v.D. acknowledges support from the Economic and Social Research Council through the Network for Integrated Behavioral Sciences (Grant: ES/K002201/1), and the Netherlands Organization for Scientific Research (Grant: 452-16-011).

Appendix. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.evolhumbehav.2021.04.002.

References

Albrecht, F., Kube, S., & Traxler, C. (2018). Cooperation and norm enforcement—The individual-level perspective. *Journal of Public Economics, 165*, 1–16. https://doi.org/10.1016/j.jpubeco.2018.06.010.

Arechar, A. A., Gaechter, S., & Molleman, L. (2018). Conducting interactive experiments online. *Experimental Economics, 21*(1), 99–131. https://doi.org/10.1007/s10683-017-9527-2.

Axelrod, R. (1986). An evolutionary approach to norms. *The American Political Science Review, 80*(4), 1095–1111. JSTOR https://doi.org/10.2307/1960858.

Balafoutas, L., Nikiforakis, N., & Rockenbach, B. (2014). Direct and indirect punishment among strangers in the field. *Proceedings of the National Academy of Sciences, 111*(45), 15924–15927.
Berkowitz, A. (2005). An overview of the social norms approach. In *Challenging the culture of college drinking: A socially situated health communication campaign (2005th ed.)*. Hampton Press.
Bicchieri, C. (2006). *The grammar of society: The nature and origins of social norms*. Cambridge University Press.
Bond, R., & Smith, P. B. (1996). Culture and conformity: A meta-analysis of studies using Asch's (1952b, 1956) line judgment task. *Psychological Bulletin, 119*(1), 111–137. https://doi.org/10.1037//0033-2909.119.1.111.
Boyd, R., Gintis, H., & Bowles, S. (2010). Coordinated punishment of defectors sustains cooperation and can proliferate when rare. *Science, 328*(5978), 617–620.
Boyd, R., Gintis, H., Bowles, S., & Richerson, P. J. (2003). The evolution of altruistic punishment. *Proceedings of the National Academy of Sciences, 100*(6), 3531–3535.
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.
Brandts, J., & Charness, G. (2011). The strategy versus the direct-response method: a first survey of experimental comparisons. *Experimental Economics, 14*(3), 375–398. https://doi.org/10.1007/s10683-011-9272-x.
Burger, J. M., Bell, H., Harvey, K., Johnson, J., Stewart, C., Dorian, K., & Swedroe, M. (2010). Nutritious or delicious? The effect of descriptive norm information on food choice. *Journal of Social and Clinical Psychology, 29*(2), 228–242. https://doi.org/10.1521/jscp.2010.29.2.228.
Burton-Chellew, M. N., El Mouden, C., & West, S. A. (2016). Conditional cooperation and confusion in public-goods experiments. *Proceedings of the National Academy of Sciences, 113*(5), 1291–1296.
Camera, G., & Casari, M. (2009). Cooperation among strangers under the shadow of the future. *American Economic Review, 99*(3), 979–1005. https://doi.org/10.1257/aer.99.3.979.
Casari, M., & Luini, L. (2009). Cooperation under alternative punishment institutions: An experiment. *Journal of Economic Behavior & Organization, 71*(2), 273–282.
Casari, M., & Luini, L. (2012). Peer punishment in teams: Expressive or instrumental choice? *Experimental Economics, 15*(2), 241–259.
Cheung, S. L. (2014). New insights into conditional cooperation and punishment from a strategy method experiment. *Experimental Economics, 17*(1), 129–153.
Chudek, M., & Henrich, J. (2010). Culture-gene coevolution, norm-psychology, and the emergence of human prosociality. *Trends in Cognitive Sciences, 15*(5), 218–226. https://doi.org/10.1016/j.tics.2011.03.003.
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. *Advances in Experimental Social Psychology, 24*, 201–234. https://doi.org/10.1016/S0065-2601(08)60330-5.
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. https://doi.org/10.1037/0022-3514.58.6.1015.
Cialdini, R. B., & Trost, M. R. (1998). Social influence: Social norms, conformity and compliance. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *The handbook of social psychology, Vols. 1 and 2 (4th ed.) (pp. 151–192)*. McGraw-Hill.
Columbus, S., & Böhm, R. (2021). Norm Shifts Under the Strategy Method. *PsyArXiv*. psyarxiv.com/5m6yt.
Crockett, M. J., Clark, L., Lieberman, M. D., Tabibnia, G., & Robbins, T. W. (2010). Impulsive choice and altruistic punishment are correlated and increase in tandem with serotonin depletion. *Emotion, 10*(6), 855–862. https://doi.org/10.1037/a0019861.
Cubitt, R. P., Drouvelis, M., & Gächter, S. (2011). Framing and free riding: Emotional responses and punishment in social dilemma games. *Experimental Economics, 14*(2), 254–272.
Egas, M., & Riedl, A. (2008). The economics of altruistic punishment and the maintenance of cooperation. *Proceedings of the Royal Society B: Biological Sciences, 275*(1637), 871–878. https://doi.org/10.1098/rspb.2007.1558.
Elster, J. (1989a). *The cement of society: A survey of social order*. Cambridge University Press.
Elster, J. (1989b). Social norms and economic theory. *Journal of Economic Perspectives, 3*(4), 99–117. https://doi.org/10.1257/jep.3.4.99.
Eriksson, K., Cownden, D., Ehn, M., & Strimling, P. (2014). "Altruistic" and "antisocial" punishers are one and the same. *Review of Behavioral Economics, 1*(3), 209–221. https://doi.org/10.1561/105.00000009.
Falk, A., Fehr, E., & Fischbacher, U. (2005). Driving forces behind informal sanctions. *Econometrica, 73*(6), 2017–2030.
Fehr, E., Fischbacher, U., & Gächter, S. (2002). Strong reciprocity, human cooperation, and the enforcement of social norms. *Human Nature, 13*(1), 1–25. https://doi.org/10.1007/s12110-002-1012-7.
Fehr, E., & Gächter, S. (2000). Cooperation and punishment in public goods experiments. *American Economic Review, 90*(4), 980–994. https://doi.org/10.1257/aer.90.4.980.
Fehr, E., & Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *The Quarterly Journal of Economics, 114*(3), 817–868. https://doi.org/10.1162/003355399556151.
Fehr, E., & Schurtenberger, I. (2018). Normative foundations of human cooperation. *Nature Human Behaviour, 2*(7), 458–468. https://doi.org/10.1038/s41562-018-0385-5.
FeldmanHall, O., Otto, A. R., & Phelps, E. A. (2018). Learning moral values: Another's desire to punish enhances one's own punitive behavior. *Journal of Experimental Psychology: General, 147*(8), 1211.
Ferraro, P. J., Vossler, C. A., & others. (2010). The source and significance of confusion in public goods experiments. *The BE Journal of Economic Analysis & Policy, 10*(1), 1–42.
Fischbacher, U., & Gachter, S. (2010). Social preferences, beliefs, and the dynamics of free riding in public good experiments. *American Economic Review, 100*(1), 541–556.
Fischbacher, U., Gächter, S., & Fehr, E. (2001). Are people conditionally cooperative? Evidence from a public goods experiment. *Economics Letters, 71*(3), 397–404. https://doi.org/10.1016/S0165-1765(01)00394-9.
Fischbacher, U., Gächter, S., & Quercia, S. (2012). The behavioral validity of the strategy method in public good experiments. *Journal of Economic Psychology, 33*(4), 897–913.
Frey, B. S., & Meier, S. (2004). Social comparisons and pro-social behavior: Testing "conditional cooperation" in a field experiment. *American Economic Review, 94*(5), 1717–1722. https://doi.org/10.1257/0002828043052187.
Gächter, S., & Herrmann, B. (2009). Reciprocity, culture and human cooperation: Previous insights and a new cross-cultural experiment. *Philosophical Transactions of the Royal Society, B: Biological Sciences, 364*(1518), 791–806.
Gächter, S., Kölle, F., & Quercia, S. (2017). Reciprocity and the tragedies of maintaining and providing the commons. *Nature Human Behaviour, 1*(9), 650–656. https://doi.org/10.1038/s41562-017-0191-5.
Gächter, S., Renner, E., & Sefton, M. (2008). The long-run benefits of punishment. *Science, 322*(5907), 1510. https://doi.org/10.1126/science.1164744.
Giamattei, M., Yahosseini, K. S., Gächter, S., & Molleman, L. (2020). LIONESS Lab: A free web-based platform for conducting interactive experiments online. *Journal of the Economic Science Association*. https://doi.org/10.1007/s40881-020-00087-0.
Gigerenzer, G. (2008). Why heuristics work. *Perspectives on Psychological Science, 3*(1), 20–29. https://doi.org/10.1111/j.1745-6916.2008.00058.x.
Gigerenzer, G. (2010). Moral satisficing: Rethinking moral behavior as bounded rationality. *Topics in Cognitive Science, 2*(3), 528–554. https://doi.org/10.1111/j.1756-8765.2010.01094.x.
Guala, F. (2012). Reciprocity: Weak or strong? What punishment experiments do (and do not) demonstrate. *The Behavioral and Brain Sciences, 35*, 1–15.
Hallsworth, M., List, J. A., Metcalfe, R. D., & Vlaev, I. (2017). The behavioralist as tax collector: Using natural field experiments to enhance tax compliance. *Journal of Public Economics, 148*, 14–31.
Hauert, C., Traulsen, A., Brandt, H., Nowak, M. A., & Sigmund, K. (2007). Via freedom to coercion: The emergence of costly punishment. *Science, 316*(5833), 1905–1907. https://doi.org/10.1126/science.1141588.
Henrich, J. (2000). Does culture matter in economic behavior? Ultimatum game bargaining among the Machiguenga of the Peruvian Amazon. *American Economic Review, 90*(4), 973–979. https://doi.org/10.1257/aer.90.4.973.
Henrich, J. (2016). *The secret of our success: How culture is driving human evolution, domesticating our species, and making us smarter*. Princeton University Press.
Henrich, J., & Boyd, R. (1998). The evolution of conformist transmission and the emergence of between-group differences. *Evolution and Human Behavior, 19*(4), 215–241.
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. *The American Economic Review, 91*(2), 73–78.
Henrich, J., Ensminger, J., McElreath, R., Barr, A., Barrett, C., Bolyanatz, A., Cardenas, J. C., Gurven, M., Gwako, E., Henrich, N., Lesorogol, C., Marlowe, F., Tracer, D., & Ziker, J. (2010). Markets, religion, community size, and the evolution of fairness and punishment. In *Science* (Vol. 327, issue 5972, pp. 1480–1484). American Association for the Advancement of Science. doi:https://doi.org/10.1126/science.1182238.
Henrich, J., McElreath, R., Barr, A., Ensminger, J., Barrett, C., Bolyanatz, A., Cardaroas, J. C., Gurven, M., Gwako, E., Henrich, N., Lesoronol, C., Marlowe, F., Tracer, D., & Ziker, J. (2006). Costly punishment across human societies. In *Science* (Vol. 312, Issue 5781, pp. 1767–1770). American Association for the Advancement of Science. doi:https://doi.org/10.1126/science.1127333.
Herrmann, B., & Thöni, C. (2009). Measuring conditional cooperation: A replication study in Russia. *Experimental Economics, 12*(1), 87–92.
Herrmann, B., Thöni, C., & Gächter, S. (2008). Antisocial punishment across societies. *Science, 319*(5868), 1362–1367. American Association for the Advancement of Science https://doi.org/10.1126/science.1153808.
Hume, D. (2003). *A treatise of human nature*. Courier Corporation.
Kamei, K. (2014). Conditional punishment. *Economics Letters, 124*(2), 199–202.
Kamei, K. (2020). Voluntary disclosure of information and cooperation in simultaneous-move economic interactions. *Journal of Economic Behavior & Organization, 171*, 234–246.
Kandori, M., Mailath, G. J., & Rob, R. (1993). Learning, mutation, and long run equilibria in games. *Econometrica, 61*(1), 29–56. https://doi.org/10.2307/2951777.
Kelley, H. H. (1971). Moral evaluation. *American Psychologist, 26*(3), 293–300. https://doi.org/10.1037/h0031276.
Keser, C., & Van Winden, F. (2000). Conditional cooperation and voluntary contributions to public goods. *The Scandinavian Journal of Economics, 102*(1), 23–39. https://doi.org/10.1111/1467-9442.00182.
Lindström, B., Jangard, S., Selbing, I., & Olsson, A. (2018). The role of a "common is moral" heuristic in the stability and change of moral norms. *Journal of Experimental Psychology: General, 147*(2), 228–242.
McGraw, K. M. (1985). Subjective probabilities and moral judgments. *Journal of Experimental Social Psychology, 21*(6), 501–518. https://doi.org/10.1016/0022-1031(85)90022-8.
Molleman, L., Kölle, F., Starmer, C., & Gächter, S. (2019). People prefer coordinated punishment in cooperative interactions. *Nature Human Behaviour*. https://doi.org/10.1038/s41562-019-0707-2.
Nikiforakis, N. (2010). Feedback, punishment and cooperation in public good experiments. *Games and Economic Behavior, 68*(2), 689–702. https://doi.org/10.1016/J.GEB.2009.09.004.

Nikiforakis, N., & Normann, H.-T. (2008). A comparative statics analysis of punishment in public-good experiments. *Experimental Economics, 11*(4), 358–369. https://doi.org/10.1007/s10683-007-9171-3.

Nolan, J. M., Schultz, P. W., Cialdini, R. B., Goldstein, N. J., & Griskevicius, V. (2008). Normative social influence is underdetected. *Personality and Social Psychology Bulletin, 34*(7), 913–923. https://doi.org/10.1177/0146167208316691.

Oosterbeek, H., Sloof, R., & van de Kuilen, G. (2004). Cultural differences in ultimatum game experiments: Evidence from a meta-analysis. *Experimental Economics, 7*(2), 171–188. https://doi.org/10.1023/B:EXEC.0000026978.14316.74.

Ostrom, E., Walker, J., & Gardner, R. (1992). Covenants with and without a sword: Self-governance is possible. *The American Political Science Review, 86*, 404–417.

Peysakhovich, A., Nowak, M. A., & Rand, D. G. (2014). Humans display a "cooperative phenotype" that is domain general and temporally stable. *Nature Communications, 5*. https://doi.org/10.1038/ncomms5939.

Raihani, N. J., & Bshary, R. (2019). Punishment: One tool, many uses. *Evolutionary Human Sciences, 1*. https://doi.org/10.1017/ehs.2019.12. e12. Cambridge Core.

Raihani, N. J., Thornton, A., & Bshary, R. (2012). Punishment and cooperation in nature. *Trends in Ecology & Evolution, 27*(5), 288–295. https://doi.org/10.1016/J.TREE.2011.12.004.

Rand, D. G., Dreber, A., Ellingsen, T., Fudenberg, D., & Nowak, M. A. (2009). Positive interactions promote public cooperation. *Science (New York, N.Y.), 325*(5945), 1272–1275. https://doi.org/10.1126/science.1177418.

Rockenbach, B., & Milinski, M. (2006). The efficient interaction of indirect reciprocity and costly punishment. *Nature, 444*(7120), 718–723. https://doi.org/10.1038/nature05229.

Roth, A. E., Prasnikar, V., Okuno-Fujiwara, M., & Zamir, S. (1991). Bargaining and market behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: An experimental study. *The American Economic Review*, 1068–1095.

Selten, R. (1967). Die strategiemethode zur erforschung des eingeschränkt rationalen verhaltens im rahmen eines oligopolexperiments. In H. Sauermann (Ed.), *Beiträge zur experimentellen Wirtschaftsforschung* (pp. 136–168). Mohr.

Sigmund, K. (2007). Punish or perish? Retaliation and collaboration among humans. *Trends in Ecology & Evolution, 22*, 593–600.

Son, J.-Y., Bhandari, A., & FeldmanHall, O. (2019). Crowdsourcing punishment: Individuals reference group preferences to inform their own punitive decisions. *Scientific Reports, 9*(1), 1–15. https://doi.org/10.1038/s41598-019-48050-2.

Szolnoki, A., & Perc, M. (2013). Effectiveness of conditional punishment for the evolution of public cooperation. *Journal of Theoretical Biology, 325*, 34–41.

Tworek, C. M., & Cimpian, A. (2016). Why do people tend to infer "ought" from "is"? The role of biases in explanation. *Psychological Science*. https://doi.org/10.1177/0956797616650875.

Weber, T. O., Weisel, O., & Gächter, S. (2018). Dispositional free riders do not free ride on punishment. *Nature Communications, 9*(1), 2390.

Welch, M. R., Xu, Y., Bjarnason, T., Petee, T., O'Donnell, P., & Magro, P. (2005). "But everybody does it…": The effects of perceptions, moral pressures, and informal sanctions on tax cheating. *Sociological Spectrum, 25*(1), 21–52. https://doi.org/10.1080/027321790500103.

Yamagishi, T. (1986). The provision of a sanctioning system as a public good. *Journal of Personality and Social Psychology, 51*(1), 110.

Young, H. P. (1993). The evolution of conventions. *Econometrica, 61*(1), 57–84. JSTOR https://doi.org/10.2307/2951778.

Young, H. P. (2015). The evolution of social norms. *Annual Review of Economics, 7*(1), 359–387. https://doi.org/10.1146/annurev-economics-080614-115322.