Biology & Philosophy (2023) 38:48
https://doi.org/10.1007/s10539-023-09936-8

The problem of opportunity

Jonathan R. Goodman[1,2]

Received: 22 May 2023 / Accepted: 22 October 2023 / Published online: 7 November 2023

Abstract
Cultural group selection theorists propose that humans evolved prosocial preferences. These claims revolve largely around the centrality of punishment in cultural groups, which helped to eliminate free riders. The purpose of this paper is to explore whether distinguishing between free-riding as an action, and free riders as entities, undermines or supports this view. I develop three individual-based models of the Prisoner's Dilemma. The first model shows that strong reciprocity removes overt freeriders from a population, and maintains a high rate of cooperation. In the second, I introduce individuals that mimic cooperative preferences, but who defect when they trick opponents into cooperating. I show that strong reciprocity is robust against this strategy, but not because individuals are replaced by strong reciprocators. Finally, I introduce a third strategy, covert mimicry, where some mimics may defect without detection. I draw attention to the problem highlighted in these models, which is that cooperation may be maintained in populations only because freeriders are not presented with the opportunity to defect. I discuss this problem in the context of cultural group selection and the human capacity for innovation, and suggest that hypotheses relying on prosocial preferences for maintaining cooperation require some revision.

Keywords Cooperation · Prisoner's dilemma · Cultural group selection · Free rider problem · Mimicry

Introduction

> Honesty is the best policy; but he who is governed by that maxim is not an honest man.
>
> - Archbishop Richard Whately

✉ Jonathan R. Goodman
jrg74@cam.ac.uk

[1] Leverhulme Centre for Human Evolutionary Studies, Cambridge, UK

[2] Darwin College, Cambridge, UK

Table 1 Overview of well-known proximate and ultimate hypotheses that aim to explain widespread human cooperation

Hypothesis	Description
Kin selection	Individuals are likely to cooperate with those with whom they share a greater proportion of genes than the average in a given population (Hamilton 1964; see also West et al. 2007)
Reciprocity	Individuals are likely to cooperate with those whom are likely to reciprocate cooperation (Trivers 1971; Axelrod and Hamilton 1981)
Indirect reciprocity	Individuals are likely to cooperate with those they witness cooperating with others (Alexander 1987; Nowak and Sigmund 1998); cooperation may also be rewarded by third parties at a later point following gossip (see Sterelny 2021)
Punishment/coercion	Individuals cooperate to avoid punishment (a partner control strategy; Frank 1998)
Partner choice	Individuals select cooperative partners known, through gossip or witnessing, to behave cooperatively (a partner choice strategy; West-Eberhard 1983; Noe and Hammerstein 1995)
Strong reciprocity	Individuals cooperate with cooperative individuals and punish non-cooperators (Gintis 2000)
Cultural group selection	Individuals inherit cultural traits promoting cooperation from others within a cultural group; cultural groups with high rates of cooperation are likely to outcompete others (Boyd and Richerson 1992, *inter alia*)
Self-domestication	Humans selected against reactive aggression over our evolutionary history, indirectly promoting qualities associated with cooperation (Wrangham 2019, *inter alia*)

The list is not exhaustive, and combinations and overlaps (e.g., strong reciprocity and cultural group selection) are common in the literature

In 2007, Dana et al. conducted a novel version of the two-player Dictator Game, where one player is given an endowment, and decides whether to give the second player any of it. Previous studies found that, against the purely economic expectation that self-interested first players should keep the whole endowment, players were likely to give a substantial portion of the endowment away, suggesting a widespread human preference for fairness and prosociality. Dana et al. (2007) found, however, that when dictator-players were given the option to pay 10% of the endowment to quietly exit the game without the second player knowing about it, just under half of players did so, keeping 90% of the endowment for themselves. Paying for the opportunity to keep more of the endowment quietly, they found, motivated a greater number of players than expected.

Several hypotheses, which have received both empirical and computational attention over the last several decades, aim to explain widespread cooperation in humans at the proximate and ultimate levels (Hamilton 1964; Trivers 1971; Alexander 1987; Frank 1988; Axelrod and Hamilton 1981; Frank 1998; Nowak and Sigmund 1998; Boyd et al. 2003; Boyd and Richerson 1992; Fehr and Fischbacher 2003; Axelrod et al. 2004; Roberts 2005; Henrich and Henrich 2007; Boehm 2017; Wrangham 2019; also see West et al. 2007); see Table 1. Prominent

among these, at least at the proximate level, is strong reciprocity (Gintis 2000; Fehr et al. 2002; Bowles and Gintis 2004), which relies on partner choice and partner control (for a discussion of this distinction, see Baumard et al. 2013) mechanisms. Given that there is almost always some economic motivation to cheat in social relationships, individuals, according to these theorists, must both choose potential partners and control existing partners with the threat of punishment (Barclay 2016; Raihani 2021).

The force of both strategies has been shown in numerous conceptual and empirical studies (for example, Gintis 2000; Fehr and Gachter 2002; Panchanathan and Boyd 2004; Wiessner 2005, although see Singh and Garfield 2022), which also tie into foundational concepts in cooperation studies, such as reputation, emotion, intelligence, and honest signalling (McElreath et al. 2003a, b; McNally et al. 2012; Számadó et al. 2021; Giardini et al. 2022). Yet a common feature of many computational experiments is that cooperation is equated with number of cooperators: in a given population, insofar as individuals cooperate in given dyadic or group-level interactions, they are considered cooperators. Research, in these cases, implicitly assumes that as the fraction of cooperative actions increases, so does the fraction of cooperators (some have also argued that "types" are stable across time, see Kurzban and Houser 2005).

The logic of cultural group selection (Boyd and Richerson 1988; Henrich 2004, 2020) goes further. Given that, as conceptual research continues to robustly show, partner control and choice strategies are effective at maintaining cooperation in human populations, it is probable that successful groups, over our evolutionary history, developed and transmitted cultural traits that promote conditional cooperation and punishment (Gintis et al. 2005; Henrich and Muthukrishna 2021). Cultural groups that effectively reduced within-group behavioural heterogeneity with relation to cooperation, according to this body of work, maximized cooperation and out-competed those with a higher proportion of free-riders in the population—those groups without effective partner choice and control mechanisms. Where cooperation flourishes, so does the cultural group, according to cultural group selection theorists.

Some authors note, however, that there remains an issue of framing: conditional cooperators are also conditional defectors (Hagen and Hammerstein 2006; Bernhard and Cushman 2022; Ibrahim 2022). In Prisoner's Dilemmas (PDs), for example, a recently noted set of tactics, known as zero-determinant strategies (Press and Dyson 2012; Hilbe et al. 2013; Stewart and Plotkin 2013), may cooperate with high frequency—but yet always ensure that the agent's payoff is at least as great as the relevant partner's. In real-world interactions, it is vague whether equivalent interactions should be labelled as prosocial, where "extortion" might be an equally apt label (Bernhard and Cushman 2022).

The ambiguity of social descriptions in agent interactions, therefore, leaves open the question of whether a prosocial strategy reflects prosocial preferences, though widespread empirical evidence suggests this assumption is nonetheless supported (Fehr and Gächter 2002; Rand et al. 2012; Henrich 2020; although see Mulder et al. 2006). Explaining the greater-than-expected degree of prosociality in public goods games and PDs across cultures has, instead, been the focus of an ever-increasing number of models in the evolutionary sciences (Henrich et al. 2005).

This focus, and related attachment to real-world findings in economic experiments, suggests that the ascriptions given to social interactions are no less important to the evolution of cooperation than are the strategies that yield the optimal payoff outcomes (Delton 2022). Coupled with the idea that individuals have more than economic reasons to cooperate, the notion that humans are instilled, genetically or culturally speaking, with prosocial preferences requires further elucidation. To determine whether the free-rider problem has been solved by cultural transmission in human groups, in other words, it is critical to understand what a free rider is—other than, in the simplest terms, an individual who defects in economic games.

The aim of this paper is to explore the notion of prosocial preferences, and their attendant consequences for cultural group selection, through an agent-based modelling paradigm that separates individual behaviours, types, and appearances.[1] My goal is to ask whether framing interactions with more specific social descriptions than those seen in many simple PD models, which explore the success of strategies only (such as in the classic work by Axelrod and Hamilton 1981), affects the interpretation of whether a high degree of cooperation is equivalent to a high degree of prosociality—that is, whether individuals who cooperate have necessarily internalized the social norms in a given culture (Sperber and Baumard 2012). High rates of cooperation, both in economic games and in wider society, may have more to do with the lack of opportunity to defect than with universal human preferences—a problem exemplified by Dana et al.'s (2007) findings.

I develop three agent-based models with random dyadic pairing that explore the success of strong reciprocity in PD interactions (see Gintis 2000). The models account for behaviour, individual type, and individual appearance as variables, which aims to represent not more complex interactions, but more complex descriptions of the interactions being modelled. I aim to capture this complexity using mimicry (see Nettle and Dunbar 1997) as a tactic for evading partner control strategies, although I do not take partner choice or reputation into account. The models progress in complexity, both in terms of strategies available and the linguistic descriptions attributable to the dyadic interactions. For example, while the strong reciprocity model can give only the simple formulation, *player x punishes player y* or *player x cooperates with player y*, the mimicry model qualifies some instances of cooperation as *mimic player x cooperates with cooperative player y*. While researchers designing these models sometimes make claims about the agents' motivations or preferences, alternate interpretations of these interactions are possible.

I discuss the distinction between cooperation and cooperators, that is cultural processes and cultural agents, in relation to previous findings in theoretical models, as well as ethnographic evidence showing higher-than-expected rates of prosociality across cultures (Henrich et al. 2005). My aim is to show that strategies that prevent defection are not unbreakable barriers to free riding, and further to suggest that, as societies grew larger, humans used more complex strategies—founded in

[1] Note that while many theoretical models promoting the idea of prosocial preferences are grounded in the strong reciprocity model, others (see West et al. 2007; Baumard et al. 2013; Guala 2012) argue that other mechanisms, for example competitive altruism, may explain empirical findings.

increasingly complex behaviours built on the use of language—to subvert cooperative norms for personal benefit. Acting honestly—or in the case of economic games, prosocially—should not, I argue, be equated with holding prosocial preferences.

Methods

General model description

(For a full model description according to the ODD standard protocol for agent-based models [Grimm et al. 2006], see the Supplement §1. Visit github.com/jonathanrgoodman/Opportunity for all source code and simulated data and the Supplement).

In the present set of models, I explored the implications of mimicry on random-pairing–based PDs. The models were designed with stepwise complexity, starting with standard strong reciprocity and graduating to mimicry-mediated covert strategies (for models of covert signalling, see Robson 1990; Smaldino et al. 2018). See Tables 2 and S1 for an overview of the evaluated parameters.

For simplicity (also see discussion), I chose to ignore elements important in previous agent-based models, including partner choice (West-Eberhard 1983; Nöe and Hammerstein 1995; Aktipis 2004; Bshary and Bergmüller 2008; Nesse 2016) reputation (Milinski et al. 2002; Nowak and Sigmund 2005; Barclay 2013), or memory (Nettle and Dunbar 1997), and instead relied entirely on partner control (sensu Gintis 2000). I also made no assumptions, except in the final model, about the degree of heterogeneity in a given population at baseline, though I do assume that behavioural mutations, both within and between generations, occur through cultural learning. This is because the first two models explore the distinction between action and preference, which I discuss further below.

In all models, at generation 1, individuals started with a random potential to reproduce (PTR) score ranging from 0–1 from a normal distribution (with a mean of 0.5 and a standard deviation of 0.2, rounded to the nearest hundredth). Individuals also started with a default action from a set of possible conditional action expressions (CAEs); these included *cooperate* and *defect* in the present models. I do not assume these variables are encoded culturally or genetically, but that a range of social and environmental factors determine these qualities. All individuals across the range of models I discuss here have both PTR and a CAE, but individuals in models 2 and 3 also have other qualities (see Table 2 and model descriptions below).

After random pairings, individuals can either cooperate (c) or defect (d), with a range of possible combinations given in Table 3; this followed standard PD models (following Axelrod and Hamilton 1981). All cooperators are assumed to be strong reciprocators who both defect against and punish defectors. The parameter values for punishing and being punished were set at -0.1 and -0.4, respectively, across the 3 models; these values were directly applied to the relevant individual's PTR (Table 4). Table 4 gives the payoff structure across models; these conformed to the standard PD game rules in the literature, where $dc > cc > dd > cd$ and $2cc > dc + cd$ (Press and Dyson 2012).

Table 2 Global and individual variables in the present set of models. For further details, see Supplement §1

Variable	Type	Description
N	Integer	Number of individuals in population
generations	Integer	Number of generations per model run
μ	Numeric (range 0–1)	Probability that an individual will adopt a new strategy on reproduction
f	Sequence	Rate of fluctuation of individual-level parameters on reproduction
μ_s	Numeric (range 0–1)	Probability that an individual will adopt or abandon a covert strategy on reproduction (model 3 only)
Conditional action expression (CAE)	Factor	Range of actions possible for an individual in dyadic interactions, includes cooperate and defect only across models
Type	Factor	Range of possible agent-types, includes honest and mimic only across models
Appearance	Factor	Range of possible agent action appearances, includes overt and silent only
Potential to reproduce (PTR)	Numeric (range 0–1)	Probability that a given individual will reproduce
Mimicry	Numeric (range 0–1)	An individual's ability to mimic cooperative tendencies, where 0 is lowest and 1 is highest
Sensitivity	Numeric (range 0–1)	An individual's ability to detect mimics, where 0 is lowest and 1 is highest

Table 3 Possible action combinations across models

	Player 2	Player 2
Player 1	Cooperate Cooperate	Defect Cooperate
Player 1	Cooperate Defect	Defect Defect

Table 4 Payoff structures to focal individual (ego) across models

Outcome	Payoff to ego
Ego cooperates, partner cooperates (cc)	0.1
Ego cooperates, partner defects (cd)	−0.1
Ego defects, partner cooperates (dc)	0.2
Ego defects, partner defects (dd)	0
Ego punishes partner for defecting	−0.1
Ego is punished for defecting	−0.4

At the end of each round of dyadic interactions and once payoffs are distributed, individuals reproduce with probability PTR and individuals in the new generation inherit parent PTR and CAE. After reproduction, PTR fluctuates by a random number in the *f* sequence (which I held steady across models at −0.4 to 0.4, in increments of 0.1), and individuals adopt a new default action with probability μ. The schematic for Model 1 is depicted in Fig. 1 using the DiagrammeR package for R (Iannone 2022).

Model 1: strong reciprocity

Model 1 is a simple version of the strong reciprocity model first developed by Gintis (2000). In this model, cooperators defect against defectors and punish them (the punishment is assumed to be costly, see Frank, 1995), directly affecting defector PTR. I assume that individuals have perfect knowledge of each other's actions, both during and after given interactions. Note that in this model, cooperators never cooperate with defectors, so the only possible payoffs in Table 4 are 0 and 0.1. In this model, type is not distinguished from strategy, so the only possible linguistic descriptions relate to individual actions, for example *player x punishes player y*. The possible interactions from Model 1 are given in Table S2.

Model 2: mimicry

In this model, I introduce two new individual-level qualities: *mimicry* and *sensitivity*, numeric variables that ranged from 0–1, where *mimicry* allowed individuals to attempt to deceive partners into cooperating while defecting, and *sensitivity* determined whether a partner could detect mimicry. To reflect this addition, I

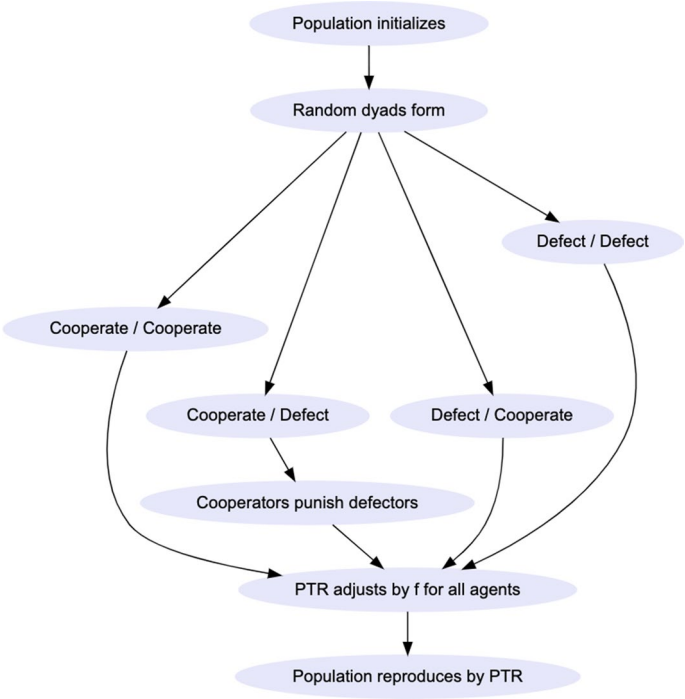

Fig. 1 Schematic of Model 1's schedule. After random dyads form, the bubbles describe the possible actions that may follow; these do not necessarily represent the types and appearances of individuals as described in Models 2 and 3. All 3 models follow this schedule

introduced another individual-level quality: *type*, which could be *mimic* or *honest* [signaller]. Individuals thus had both strategy (*CAE*; cooperate/defect) and *type*.

Individuals each had random *mimicry* and *sensitivity* scores at generation 0 from a normal distribution; as with PTR, *mimicry* and *sensitivity* fluctuated by *f* after dyadic interactions. Mimics defaulted to defect and attempted to trick honest signallers into cooperating, and defected when mimicry was successful; accordingly, *type* allows for a more complex linguistic description than in Model 1, for example *mimic x defects against honest signaller y*.

The mimicry model was identical to the strong reciprocity model in all respects except attempts at mimicry. Honest cooperators and defectors did not make use of mimicry, but only sensitivity; mimics used both mimicry and sensitivity. On random pairing, a defector switches to cooperate if the relevant partner has at least twice the focal individual's mimicry score ($mimicry_{focal} \leq 2 \times sensitivity_{partner}$). This behaviour was stored as an instance of cooperation in the analysis; after the interaction, the mimic switched back to their default defect strategy.

Otherwise, the focal individual defects and the payoff structure in Table 4 holds, differing from Model 1 in that mimics with $mimicry_{focal} > 2 \times sensitivity_{partner}$ deceive partners into cooperating, receiving the maximum payoff from

Table 3, while the cheated cooperator receives the sucker's payoff. Honest defectors could not receive the maximum payoff, as in Model 1.

Finally, as with Model 1, individuals switched strategy with probability μ. In addition, however, individuals also switched type with probability μ using an unrelated probabilistic function; switching type did not affect individual strategy. (For example, an honest cooperator switching type did not necessarily switch strategy. It was therefore possible to switch from an honest cooperator to a mimic cooperator in order that mutation of type did not add significantly to the change in frequency of cooperators). The possible interactions from Model 2 are given in Table S3.

Model 3: silent defectors

Finally, in the silent defector model, I introduce a third individual-level variable: *appearance*. The default appearance for all individuals was overt, meaning that any strategy-type combination (honest cooperator, honest defector, mimic cooperator, mimic defector) was overt, which in these models means "visible to all partners."

Again, the model was identical to Model 2, except that, at each reproduction phase, mimics mutated to the silent appearance with probability μ_s (and those with the silent appearance in generations after 1 had the same probability of mutating back to overt appearance). Mutation of appearance did not affect other individual-level attributes, such as strategy. *Appearance* made possible linguistic descriptions still more complex than in Model 2, for example *mimic x silently defects against honest signaller y*.

Mimics who "discovered" or mutated the silent signalling strategy faked the cooperation strategy unless $mimicry_{focal} \leq sensitivity_{partner}$, in which case they, as in Model 2, switched to cooperate for the duration of the dyadic pairing. As with Model 2, individuals who switched to cooperate were recorded as cooperating in the relevant generation.

Where $mimicry_{focal} > sensitivity_{partner}$ among those with the silent mimicry appearance, ego defected and gained the maximum payoff where partners cooperated. The defection was, however, not detectable, which I propose was due to the mimic using a novel strategy to avoid the partner having knowledge of the mimic's identity—a quality I discuss further below. Mimics with high mimicry scores who mutated to the silent appearance, therefore, defected without incurring punishment. Notably, furthermore, these individuals punished defectors in dyadic interactions, in keeping with behaviours proposed in the literature around the second-order free-rider problem (Heckathorn 1989; Panchanathan and Boyd 2004; see also Matthew 2017). The possible interactions in Model 3 are given in Table S4.

Analyses

I ran a proof-of-concept initial PD model without strong reciprocity to show the model's general functionality. I then ran each model for 100,000 generations (except for the basic PD model and for Model 1, which each ran for ~ 3000 generations, with a probability function for randomly ending the model loop) with set

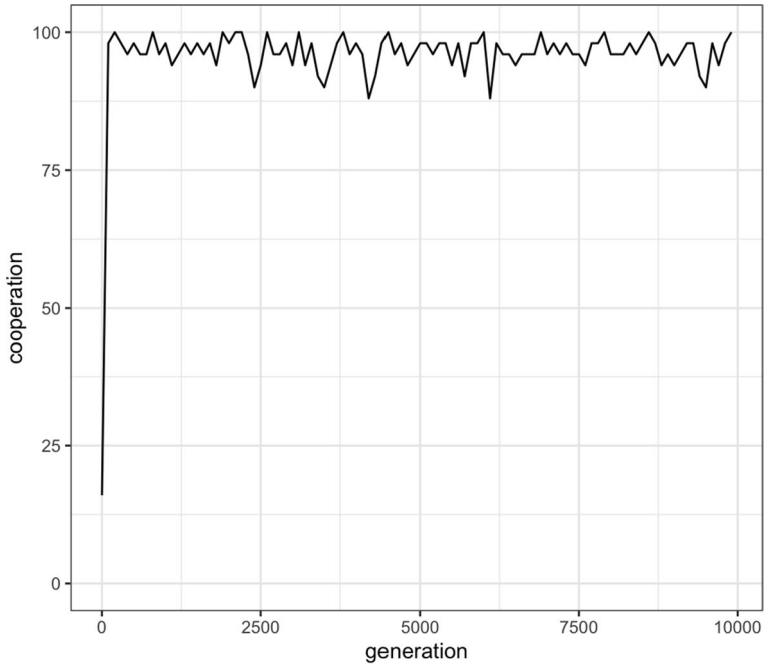

Fig. 2 Percentage of cooperation in the strong reciprocity model with the above parameters; the model ran for 10,000 generations (values taken every 100 generations for viewability). Cooperation remained nearly 100% despite a high mutation rate of 0.1. Parameters: N = 100; generations = 10e4; μ = 0.1; ∂ = 0.0001

global parameters, and calculated the frequency of cooperation as a percentage over the generations, as well as fluctuations around individual strategies (cooperate vs defect), types (mimic vs honest), and appearance (overt vs silent). All scripts were written using basic R functions in R Studio (R Core Team 2022); results were represented using the ggplot2 (Wickham et al. 2021) package for R.

Results

Strong reciprocity model (Model 1)

I initially ran a proof-of-concept PD model without conditional cooperation or punishment; the results are in Supplement §2 figures S1 and S2. As expected, and following early research into the evolution of cooperation (see Axelrod and Hamilton 1981), defection was consistently the best strategy assuming no memory of repeated interaction, with defectors quickly replacing cooperators, even with a mutation rate of 0.1.

Figure 2 gives the results from the strong reciprocity model designed to follow the basic principles as set out by Gintis (2000). With identical parameters to those in the initial PD model without strong reciprocity (N = 100; generations = 10e4;

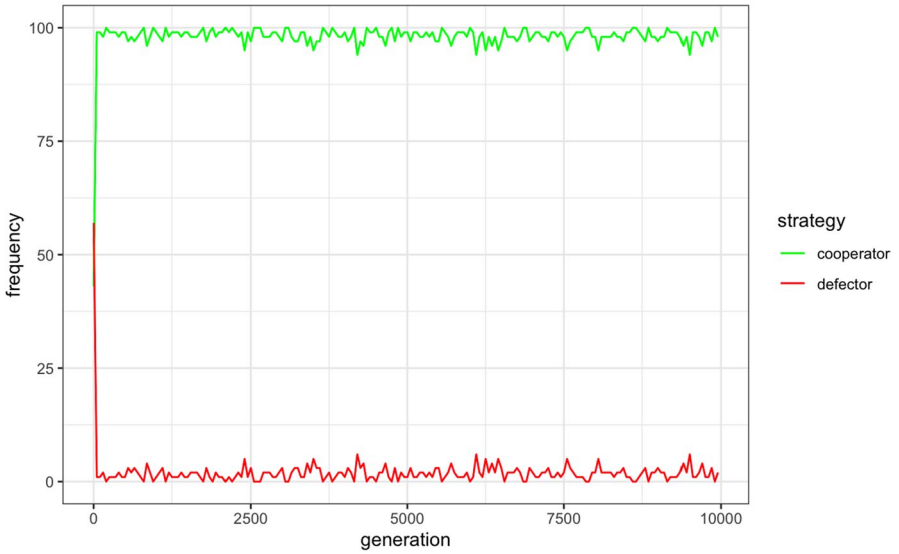

Fig. 3 Frequency of cooperators (green) and defectors (red) in the strong reciprocity model with conditional cooperation and punishment over 10,000 generations (values taken once every 100 generations for viewability). The number of cooperators in the population corresponds to the percentage of cooperative actions in the population. Parameters: N = 100; generations = 10e4; $\mu = 0.1$; $\partial = 0.0001$

$\mu = 0.1$; $\partial = 0.0001$), cooperation quickly rose to fixation, and held near 100% despite a high mutation rate.

I note in particular that the relative frequency of cooperators in the population is almost identical to the percentage of cooperation (Fig. 3). There is, otherwise put, no distinction in this model between individuals who cooperate and cooperators, and individuals who defect and defectors.

Mimicry model (Model 2)

The mimicry model was identical to the strong reciprocity model, except that agents could differ by type (honest, mimic) as well as by action (cooperate, defect). Individuals also had *sensitivity* (0–1) and *mimicry* (0–1) scores to simulate the competition between disguise as a cooperator and detection of mimicry (see Methods).

Figures 4 and 5 give contrasting results to those in the strong reciprocity model, with nearly identical parameters (N = 100; generations = 10e4; $\mu = 0.01$; $\partial = 0$). Over the 100,000 generations, the percentage of cooperative actions is almost always higher than 50% (median, 82% [range 44–99%]; Fig. 4), though the frequency of defectors (honest or mimic) is always higher than the frequency of cooperators (Fig. 5).

Figure 6, lastly, gives the frequency of mimics and honest signallers over the 100,000 generations. Neither mimics nor cooperators consistently remain at a higher frequency where $\mu = 0.01$).

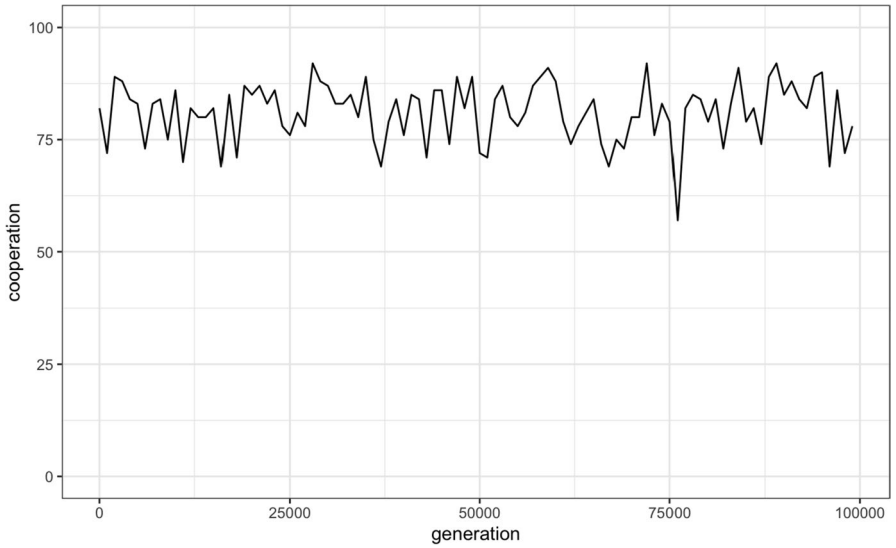

Fig. 4 Percentage of cooperation over 100,000 generations (values taken every 1000 generations for viewability) in Model 2. Median cooperation over the whole period is 82% (range 44–99%). Parameters: N = 100; generations = 10e4; μ = 0.01; $\partial = 0$

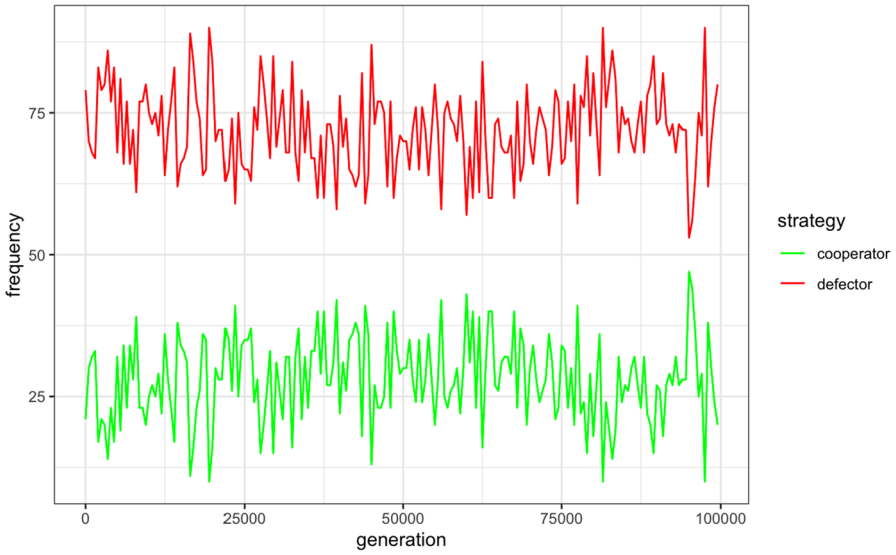

Fig. 5 Frequency of cooperators (green) and defectors (red) over 100,000 generations (values taken every 1000 generations) in Model 2. Defectors (honest and mimic, grouped together) generally have a higher frequency in the population throughout the run. Parameters: N = 100; generations = 10e4; μ = 0.01; $\partial = 0$

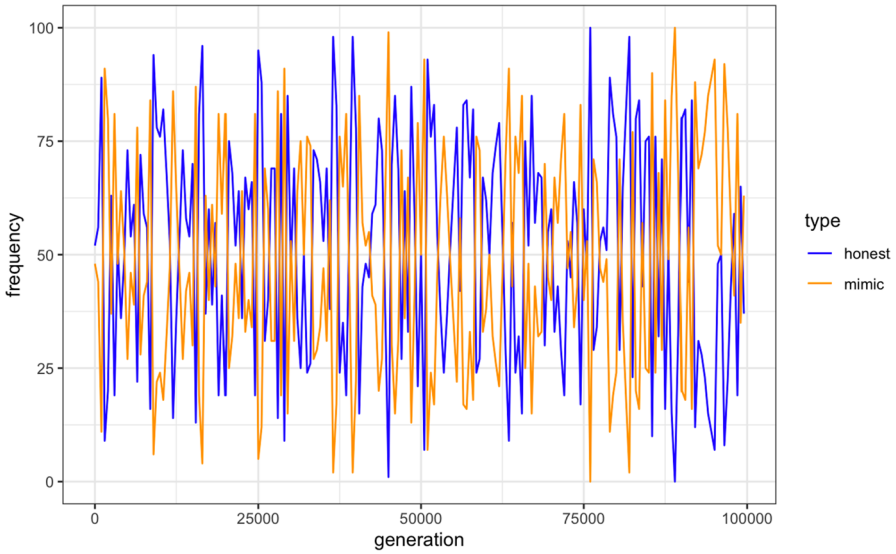

Fig. 6 Frequency of honest signallers (blue) and mimics (orange) over 100,000 generations (values taken every 1000 generations) in Model 2. Honest signallers may be defectors or cooperators by default; mimics default to defect. Neither type is consistently more frequent over the model's run. Parameters: N = 100; generations = 10e4; $\mu = 0.01$; $\partial = 0$

Covert mimicry (Model 3)

The covert mimicry model is identical to the mimicry model, except that individuals have default behaviours (cooperate, defect), types (honest, mimic) and appearances (overt, covert; see Methods). Unlike with default behaviours and types, however, the population is not initially mixed, and all agents have the overt appearance at baseline. I also added a new variable, μ_s, the probability that an overt appearance mutates to covert, and vice versa.

Figure 7 gives the percentage of cooperation vs the number of cooperators over 100,000 generations with identical parameters to the mimicry model (N = 100; generations = 10e4; $\mu = 0.01$; $\partial = 0$); I also set μ_s at 0.0001). On first mutation to covert appearance, the percentage of cooperation drops from centering above 75% to centering above 50%; this trend continues throughout the remaining generations (median cooperation, 60% [range 27–97%]).

Once the first mutation to the covert appearance occurs, furthermore, individuals with that appearance quickly go into fixation, with occasional mutations back to the overt appearance (Fig. 8). This has meaningful impacts on the default strategies and types of individuals in the population, even though cooperation remains largely the majority behaviour. Figure 8 includes the frequencies of defector strategy, mimic type, and silent appearance over 100,000 generations in Model 3. See the online supplement for further plots contrasting strategies, types, and appearances.

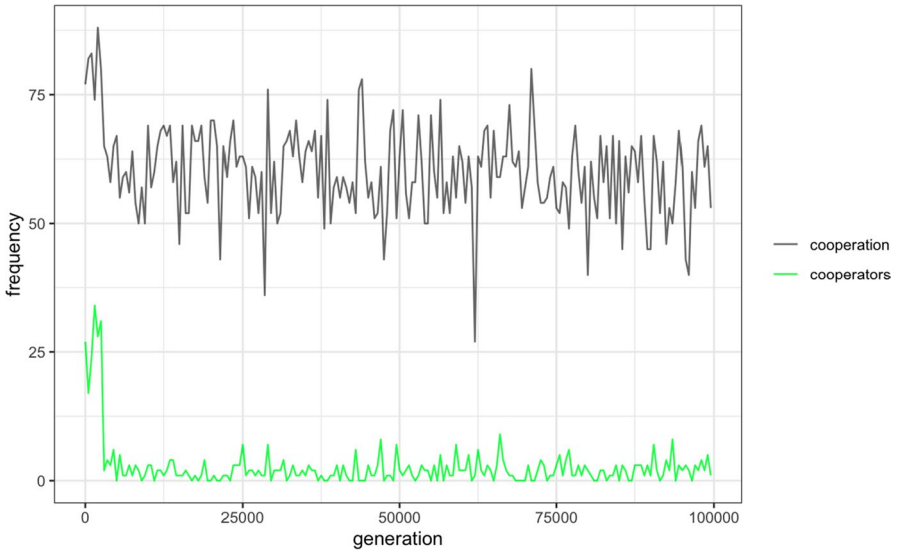

Fig. 7 Percentage of cooperation vs cooperators over 100,000 generations with the above parameters in the covert mimicry model (values taken once every 1000 generations) in Model 3. Cooperation drops from centering above 75% to centering above 50% after first mutation to the covert appearance (median cooperation, 60% [range 27–97%]); the number of cooperators, however, is consistently lower. Parameters: N = 100; generations = 10e4; $\mu = 0.01$; $\mu_s = 0.0001$; $\partial = 0$

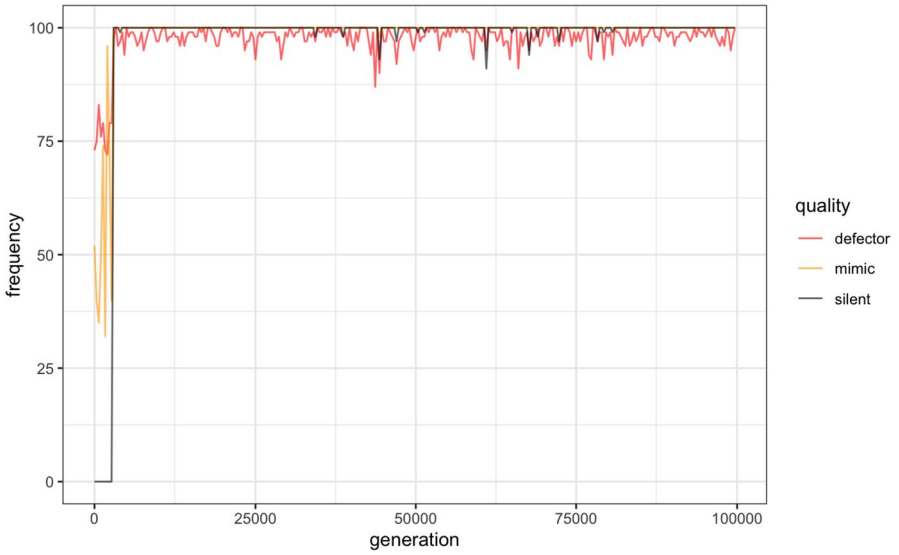

Fig. 8 The frequencies of defector strategy, mimic type, and silent appearance (grouped as individual "qualities") in the population over 100,000 generations (values taken once every 1000 generations) in Model 3. After initial mutation to the covert appearance, this trait largely remains in fixation, and the defector strategy and mimic type also increase to nearly 100% after silent mimics enter the population, though there is some fluctuation. This finding is of particular note given the high rate of cooperation still noted in the population (Fig. 7). Parameters: N = 100; generations = 10e4; $\mu = 0.01$; $\mu_s = 0.0001$; $\partial = 0$

Discussion

The aims of this paper are two-fold: to show, in line with previous suggestions in theoretical discussions (Trivers 1971), that the percentage of cooperative behaviours and the number of cooperators in a population are not identical, and that strategies that allow agents to evade detection do not markedly decrease the percentage of cooperative behaviours, but regardless spread in the population. I also suggest that these findings call into question the force of model-derived hypotheses from strong reciprocity theorists. The possibility of invasion by unknown strategies, such as silent defection, which, by definition, simple models cannot account for, allows for the illusion of high rates of cooperation without a high frequency of cooperators. Cooperation is phenotypic, and while two strategies may be phenotypically identical, we cannot speculate about the underlying motivations driving them. In this section, I discuss the implications of each of these points within the context of previous work.

Cooperation versus cooperators

In his classic paper on reciprocity, Trivers (1971) noted that individuals are unlikely to be cooperators or free riders, but to switch between strategies depending on the context. The question of whether individuals belong to the class "cooperator" or "free rider" has been put in previous work (for example, McElreath et al. 2003a, b), though it remains largely unexplored in models simulating the circumstances under which cooperation is likely to evolve.

Classic strategies (Tit-for-Tat, see Axelrod and Hamilton 1981; "forgiving", see Fudenberg et al. 2012; GRIM, Friedman 1971; Axelrod 2000) and newer strategies, such as zero-determinant strategies (Press and Dyson 2012), cooperate or defect conditionally, though in the case of Tit-for-Tat, the default behaviour on first interaction is cooperate—earning this strategy the label "cooperator." My aim with the paradigm presented in the above set of models is, firstly, to question the justification of this label.

The initial model gives us only the information that agent x did or did not cooperate with y, and vice versa. This leaves open the possibility that x would have behaved differently, had the circumstances been different, and moreover equates the agent's action with any underlying motivations that led to the action (see Hagen and Hammerstein 2006; Sperber and Baumard 2012; Casey et al. 2021). We should not, in other words, assume that a high frequency of cooperation implies a high number of cooperators in a given population of agents—the circumstances are too vague to tell us more than that, under the circumstances given, cooperation is likely to evolve and become common in a population (as, for example, is seen in many models with Tit-for-Tat and punishment strategies).

Evading detection

The design of the above models aims to make clear that "*mimic x cooperates with honest signaller y*" appears to be the same interaction as "*player x cooperates with player y*." Yet the extra linguistic information given in the former formulation identifies the mimic as an individual who cooperates only because of the circumstances, not because of its status as a cooperator. Thus individuals whose default behaviour would be to defect cooperate not necessarily because of prosocial preferences, but possibly because they perceive that the threat of detection is too high to risk exposure to punishment.

When, as in this model, we introduce the possibility of hiding defection, these individuals continue to cooperate at a high rate, though strong reciprocators—at least, in their simplest form—are almost entirely replaced in the population, and the covert strategy goes into fixation shortly after first mutation to it. This emulates the results of Nettle and Dunbar (1997), which explored the impact of a mimic strategy in a PD, though these authors found that an increasingly complex signal, coupled with high memory capacity, can prevent mimics from overwhelming a population (though also see Wiseman and Yilankaya 2001; this may relate to the importance of trust based on signals in determining whether to cooperate with a partner [Han et al. 2021]). In the present model, however, the covert mimicry strategy may prevent anti-defection strategies altogether, either through cooperation with highly sensitive individuals or through defection without detection.

An interesting question will be the degree to which opportunistic individuals can avoid detection in their relevant cultures, and whether their strategies reflect an arms race between mimicry and mimicry sensitivity over cultural evolutionary time. While Sperber and Baumard (2012) suggest that Machiavellian-like social qualities (following Humphrey 1976; Byrne and Whiten 1988) may be frequency-dependent in a given society—otherwise free-riding would quickly go into fixation and be open to invasion by cooperators—it is unlikely that detection evasion would be an effective long-term strategy in small-scale societies (for a contemporary example, see Wiessner 2005). As humans probably evolved living in groups where individuals had an average of no more than 150 close connections (Dunbar 1993; but see Lindenfors et al. 2021, for a recent critique of Dunbar's calculations), the circumstances under which one could mask one's identity, or to otherwise defect without detection, would have been limited, save for some of the circumstances Trivers (1971) notes, such as refusing help to someone likely to die.

As groups grew in size, however, the opportunities for defection without detection would have been likely to increase (Moreau 2020). Human groups began and continue to rely on concentric circles of connections; these wider "cognitive groups" allow for greater opportunities for subtle cases of deceit and defection by outsiders. While many researchers accept the notion that punishment and sanctions against norm violators kept defection in check (Fehr and Gachter 2002; Panchanathan and Boyd 2004), it is possible that, following Sperber and Baumard's (2012) prediction, a small number of such violators might be able to successfully endure. Strategies might include masquerading as an in-group member by faking tags (as distinguished from signals, see Axelrod et al. 2004) associated with membership (for a general

model of tag-based cooperation, see Riolo et al. 2001; see also McElreath et al. 2003a, b; Traulsen and Nowak 2007; Cohen 2012; Cohen and Haun 2013; Moya and Boyd 2016; Bell and Paegle 2021; see also Goodman et al. 2023), or moving quickly between inhabited areas, placing costs on locals and leaving before others notice one's defections. Moreover, most previous models do not explore the cost of checking whether one has been cheated (Han et al. 2021). Furthermore, there is nothing to suggest, either in models or in studies of small-scale cultures, that should circumstances change, new behavioural mutations would necessarily not invade. Wrangham's (2021; see also Hare 2017) position that ancient humans selected against reactive aggression as part of the self-domestication process does not, as an example, suggest that, as societies grew larger, proactive aggression and Machiavellian strategies could not emerge that subverted within-group cooperation.

To counter these risks, growing communities were likely to invest in mechanisms such as policing to enforce social norms around cooperation (Tullock 2004, *inter alia*). These investments and their consequent mechanisms reflect an arms race between individuals who aim to subvert cooperative systems through mimicry and those who cooperate honestly. Yet as social systems grew (and continue to grow) in complexity, improved mechanisms for detecting cheaters were (and are) required, in an analogous way to how immune systems evolve to thwart disease (sensu Aktipis 2020; see also Goodman and Ewald 2021). The consequence is that, across large-scale societies, we do not see the virtual ubiquity of cooperation predicted by strong reciprocity models. Novel methods for defection and exploitation also emerge regularly. One recent example in Western societies is affinity fraud, where individuals use signals reflecting shared social identity to garner trust and then exploit victims, usually for financial reasons (Blois and Ryan 2013). These and analogous dishonest signals of cooperative intent are, however, likely to become costlier to develop and enact as cultural groups improve detection strategies.[2]

Empirical tests may explore the validity of this set of points. In a public goods game, for example, it may be that individuals who appear to be following norms—giving some, but not too much (see Barclay and Willer 2007)—may end with a greater number of resources insofar as they defect only in the final round. Analogous points might be true in ultimatum, dictator and PD games—and may help to explain the finding that individuals who score high psychopathy tests tend to do well in laboratory experiments of PD games (Mokros et al. 2008). Moreover, frequency-dependent opportunism may explain why, despite evidence for hypotheses around self-domestication and cultural group selection, psychopathy continues to pervade society.

Moreover, while researchers, particularly in evolutionary psychology, suggest that humans evolved a cheater detection module (Cosmides and Tooby 1992; see also Verplaetse et al. 2007), the models presented here suggest it is unlikely that any module is effective enough to thwart all free riders in a given culture. Previous research suggests, for example, that mimicry of honest human social signals is detected in about two-thirds of cases (laughter, Bryant and Aktipis 2014; lies,

[2] I thank one of the reviewers of an older version of this article for bringing up this point.

Fonseca and Peters 2021; cooperative intent, Verplaetse et al. 2007; accents of language, Tate, 1979, Goodman et al. 2021). Yet where honest signallers and mimics (and overt and silent signallers) interact over generations, a two-thirds average detection rate suggests an arms race between these types and appearances, while maintaining cooperation at a high level (see Supplement §3).

Cultural evolution and internalization

One possible objection to the claims I make here is that punishment strategies evolved into norms, which spread across societies in our evolutionary past (Boyd and Richerson 2002; Boyd et al. 2011; for a study of relevance to religion see, for example, Atran and Henrich 2010). Cultural groups in which individuals behaved similarly—cooperating within the culture and punishing sanction violators—were likely to outcompete groups without effective cooperation-enforcing norms (Henrich and Henrich 2007). Furthermore, to the extent that individuals within successful cultures internalized the norms enforcing cooperation, which may entail forming prosocial preferences (Henrich and Muthukrishna 2021), group success–maximizing behavioural trends were likely to spread.

I am not claiming, however, that groups with opportunistic free riders who mimic cooperators are likely to be less successful. It may even be that, following some claims given in the self-domestication hypothesis (e.g., Wrangham 2021, *inter alia*), ancient humans selected for, not just cooperative tendencies, but intelligence for navigating complex social situations (the social brain hypothesis, Humphrey 1976; Gavrilets and Vose 2006; Shultz and Dunbar 2007; see also Dunbar and Shultz 2017). If true, it is possible that successful cultural groups do not only consist of cooperative individuals, but individuals capable of steering society—and consequently, themselves—to success through effective strategizing that may involve Machiavellian strategizing (Humphrey 1976). Internalization is not a valid objection, as the models presented here make no claim about beliefs, but rather the possibility of disguised defection. Previous work shows, for example, that individuals are effective at justifying transgressions (see Mazar et al. 2008), and self-deception may help reduce the cognitive load of free riding when opportunities are presented (e.g., Trivers 2000).

A second objection is that disguised mimics are likely to overwhelm a population of free riders, leaving them open to invasion by strong reciprocators through cultural evolutionary processes. For example, given that humans have an evolved bias to copy successful individuals, or to conform to behaviours around them (Boyd and Richerson 2002; Henrich and Henrich 2007), free riders are likely to be copied until the behaviour reaches fixation. While this is a likely outcome where appearances are overt, the silent mimicry strategy cannot necessarily be copied—the successful individuals will appear to be cooperators even if they defect. This prevents the behaviour spreading through a conformity or prestige bias, and maintains silent defection as a frequency-dependent strategy. Insofar as the silent, opportunistic defector does not teach its behaviour to others in the next generation (or in a genetic model, does not

reproduce more than others in the population), the strategy will not overwhelm a population of cooperators—a simulation worth exploring in future models.

Finally, one may object that these findings relate only to cooperation in dyads, while major issues in human cooperation are more likely to concern large-scale social dilemmas. This is an invalid objection for two reasons. First, dyadic meetings between strangers—that is, where agents have little to no information about their potential cooperative partners—remain an evolutionary problem from a game-theoretical point of view. Second, and more importantly, covert mimicry is an equally concerning strategy in large-scale social dilemmas, where free riders may cheat, potentially with less risk of detection than in dyadic pairings. Further research, both modelling and empirical, may confirm this suggestion.

Language and the Ring of Gyges effect

A further strategy for masking uncooperative behaviour is language: insofar as individuals can effectively justify norm transgressions, they can avoid sanctions (relying on the distinction between a rule's wording and judgments about it, see Rawls 1955). Language is therefore an effective tool for masking opportunistic defection, and presents strategic individuals with opportunities for defection in complex social scenarios.[3] This may involve lying—or just effective rhetoric—but there is no established reason to suppose that either strategy is likely to be out-competed by individuals who internalize norms of strong reciprocity. The linguistic descriptions attending the present set of models attest to this; the simple descriptions of strategies in Model 1 give far less information than do the types and appearances of Models 2 and 3, respectively. There is, however, no reason to assume that, in any real-world scenario, attributing adjectives and adverbs to social interactions will not colour the actions of individuals in ways that elicit moral judgment. Saying *Donald cooperated with Boris* does not imply that Donald is a cooperative person, or that he does not also silent defect against Boris, should the opportunity arise.

A further but related problem may be found in Plato's *Republic*, where Gyges of Lydia, a shepherd, found a ring in a cave that allowed him to become invisible. Gyges used his newfound power to murder the king, marry the queen, and install himself as ruler. While invisibility is not possible, language allows individuals to hide their defections from others, saying, as in Model 3, that a player silently defects assumes perfect knowledge. And yet if Donald were to silently defect against Boris, by definition, no one would know about it. Language creates, in other words, a Ring of Gyges effect through which strategic humans hide their exploitative actions.

[3] Language can, of course, also be used to enforce cooperation through threats of gossip and denunciation.

Consequences and future directions

Taken together, the findings from the models given here, coupled with the human ability to navigate social circumstances effectively using language, suggest that models in which individuals are presented only with a few options cannot account for strategies for defection while avoiding punishment. Following Dennett (1988), who notes, following the philosopher Karl Popper, we should assume that the human ability to innovate—creating previously unknown strategies in social competitions—is problematic for models that account only for a set of established strategies in games like the PD. Future work should, following the results of Dana et al. (2007), try to address this by giving, in laboratory conditions, only a subset of individuals an opportunity for hidden defection, and establishing whether individuals are likely to take the opportunity. Similarly, these findings should make ethnographic researchers more cautious when making claims about human prosocial preferences, such as inequity aversion, following experiments using economic games (Fehr and Schmidt 1999). Humans across cultures may not act in these games entirely in their economic self-interest, but it does not follow that these behaviours reflect prosociality.

The problem of opportunity is that individuals may not defect in cooperation-related games not because of prosocial preferences, but because they do not have the opportunity to do so. This does not undermine the cultural group selection arguments based on internalized norms and prosocial preferences (e.g., Henrich 2020), but rather suggests that such internalization and prosociality may not be as widespread as supposed in culturally successful groups (see Burton-Chellew and West 2013). Some individuals may only mimic the behaviours of devout actors (Singh and Hoffman 2021; see also Atran and Henrich 2010, for a discussion of the concept of devout actors), and defect when they have the chance, or think of a new way to do so.

The models presented may be explored with other factors in mind. First, unlike with other PD models representing dyadic interactions, I do not account for genetic (or cultural) flow; individuals cannot move between islands (as, for example, in Leimar and Hammerstein 2001). This is an unrealistic assumption, but for simplicity I do not consider the alternative here. Similarly, and more importantly, I ignore the influence of indirect reciprocity, reputation, and biological markets. While I recognize these are foundational elements in non-kinship-based cooperation models, I again do not consider them here for the sake of simplicity. In both cases, furthermore, I suggest that hidden strategies are no less likely to evolve if accounted for in the models, but rather the manifestation of hidden defection will be different. Future work should evaluate this assumption.

Conclusion

In this paper, I have suggested that the possibility of hidden mimicry presents a problem for theories of cultural group selection relying on strong reciprocity. While prosociality is essential for success in between-group competition and to a functional society, individuals who cooperate only insofar as it benefits them—and

moreover, who defect when presented with the opportunity—may thrive within a population without reducing cooperation below 50% in a simulation of agents. Individuals may therefore behave prosocially without being prosocial—an element of human psychology that researchers cannot account for in agent-based models. While cooperation, like honesty, may be the best policy, we should not assume that individuals who cooperate are themselves cooperators. The complexity of language, and the uniquely human ability to represent oneself differently in different social situations, make the possibility of undetected defection real in everyday scenarios. With growing civil unrest across the world, and with a growing number of political leaders who continue to undermine cooperative norms, future work in the social sciences should look to ways to promote cooperation, rather than to assume its pervasiveness.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10539-023-09936-8.

Acknowledgements I am indebted to Robert Foley, Daniel Nettle, Nikhil Chaudhary, David Lahti, Adam Hunt, and two anonymous reviewers for helpful comments on previous drafts of this essay.

Declarations

Conflict of interest I have no relevant conflicts to disclose.

References

Aktipis CA (2004) Know when to walk away: contingent movement and the evolution of cooperation. J Theor Biol 231(2):249–260. https://doi.org/10.1016/j.jtbi.2004.06.020

Aktipis CA (2020) The cheating cell. Princeton University Press

Alexander R (1987) The biology of moral systems

Atran S, Henrich J (2010) The evolution of religion: how cognitive by-products, adaptive learning heuristics, ritual displays, and group competition generate deep commitments to prosocial religions. Biol Theory 5(1):18–30. https://doi.org/10.1162/BIOT_a_00018

Axelrod R (2000) On six advances in cooperation theory. Anal Kritik 22(1):130–151. https://doi.org/10.1515/auk-2000-0107

Axelrod R, Hamilton WD (1981) The evolution of cooperation. Science (new York, n.y.) 211(4489):1390–1396. https://doi.org/10.1126/science.7466396

Axelrod R, Hammond RA, Grafen A (2004) Altruism via kin-selection strategies that rely on arbitrary tags with which they coevolve. Evol Int J Org Evol 58(8):1833–1838. https://doi.org/10.1111/j.0014-3820.2004.tb00465.x

Barclay P (2013) Strategies for cooperation in biological markets, especially for humans. Evol Hum Behav 34(3):164–175. https://doi.org/10.1016/j.evolhumbehav.2013.02.002

Barclay P (2016) Biological markets and the effects of partner choice on cooperation and friendship. Curr Opin Psychol 7:33–38. https://doi.org/10.1016/j.copsyc.2015.07.012

Barclay P, Willer R (2007) Partner choice creates competitive altruism in humans. Proc R Soc B Biol Sci 274(1610):749–753. https://doi.org/10.1098/rspb.2006.0209
Baumard N, André J-B, Sperber D (2013) A mutualistic approach to morality: the evolution of fairness by partner choice. Behav Brain Sci 36(1):59–78. https://doi.org/10.1017/S0140525X11002202
Bell AV, Paegle A (2021) Ethnic markers and how to find them. Hum Nat 32(2):470–481. https://doi.org/10.1007/s12110-021-09401-z
Bernhard RM, Cushman F (2022) Extortion, intuition, and the dark side of reciprocity. Cognition 228:105215. https://doi.org/10.1016/j.cognition.2022.105215
Blois K, Ryan A (2013) Affinity fraud and trust within financial markets. J Financ Crime 20(2):186–202. https://doi.org/10.1108/13590791311322364
Boehm C (2017) Moral origins. https://www.basicbooks.com/titles/christopher-boehm/moral-origins/9780465029198/
Bowles S, Gintis H (2004) The evolution of strong reciprocity: cooperation in heterogeneous populations. Theor Popul Biol 65(1):17–28. https://doi.org/10.1016/j.tpb.2003.07.001
Boyd R, Richerson PJ (1988) Culture and the evolutionary process, 2nd edn. University of Chicago Press
Boyd R, Richerson PJ (1992) Punishment allows the evolution of cooperation (or anything else) in sizable groups. Ethol Sociobiol 13(3):171–195. https://doi.org/10.1016/0162-3095(92)90032-Y
Boyd R, Richerson PJ (2002) Group beneficial norms can spread rapidly in a structured population. J Theor Biol 215(3):287–296. https://doi.org/10.1006/jtbi.2001.2515
Boyd R, Gintis H, Bowles S, Richerson PJ (2003) The evolution of altruistic punishment. Proc Natl Acad Sci USA 100(6):3531–3535. https://doi.org/10.1073/pnas.0630443100
Boyd R, Richerson PJ, Henrich J (2011) Rapid cultural adaptation can facilitate the evolution of large-scale cooperation. Behav Ecol Sociobiol 65(3):431–444. https://doi.org/10.1007/s00265-010-1100-3
Bryant GA, Aktipis CA (2014) The animal nature of spontaneous human laughter. Evol Hum Behav 35(4):327–335. https://doi.org/10.1016/j.evolhumbehav.2014.03.003
Bshary R, Bergmüller R (2008) Distinguishing four fundamental approaches to the evolution of helping. J Evol Biol 21(2):405–420. https://doi.org/10.1111/j.1420-9101.2007.01482.x
Burton-Chellew MN, West SA (2013) Prosocial preferences do not explain human cooperation in public-goods games. Proc Natl Acad Sci USA 110(1):216–221. https://doi.org/10.1073/pnas.1210960110
Byrne RW, Whiten A (eds) (1988) Machiavellian intelligence: social expertise and the evolution of intellect in monkeys, apes, and humans. Clarendon Press/Oxford University Press, pp xiv, 413
Casey W, Massey SE, Mishra B (2020) How signalling games explain mimicry at many levels: from viral epidemiology to human sociology. J R Soc Interface 18(175):20200689. https://doi.org/10.1098/rsif.2020.0689
Cohen E (2012) The evolution of tag-based cooperation in humans: the case for accent. Curr Anthropol 53(5):588–616. https://doi.org/10.1086/667654
Cohen E, Haun D (2013) The development of tag-based cooperation via a socially acquired trait. Evol Hum Behav 34(3):230–235. https://doi.org/10.1016/j.evolhumbehav.2013.02.001
Cosmides L, Tooby J (1992) Cognitive adaptations for social exchange. In: The adapted mind: evolutionary psychology and the generation of culture. Oxford University Press, pp 163–228
Dana J, Weber RA, Kuang JX (2007) Exploiting moral wiggle room: experiments demonstrating an illusory preference for fairness. Econ Theor 33:67–80. https://doi.org/10.1007/s00199-006-0153-z
Delton AW (2022) Are we there yet? Every computational theory needs a few black boxes, including theories about groups. Behav Brain Sci 45:e103. https://doi.org/10.1017/S0140525X21001217
Dennett D (1984) Cognitive wheels: the frame problem of AI. In: Hookway C (ed) Minds, machines and evolution. Cambridge University Press
Dunbar RIM (1993) Coevolution of neocortical size, group size and language in humans. Behav Brain Sci 16(4):681–694. https://doi.org/10.1017/S0140525X00032325
Dunbar RIM, Shultz S (2017) Why are there so many explanations for primate brain evolution? Philos Trans R Soc B Biol Sci 372(1727):20160244. https://doi.org/10.1098/rstb.2016.0244
Fehr E, Fischbacher U (2003) The nature of human altruism. Nature 425(6960):785–791. https://doi.org/10.1038/nature02043
Fehr E, Gächter S (2002) Altruistic punishment in humans. Nature 415(6868):137–140. https://doi.org/10.1038/415137a
Fehr E, Schmidt KM (1999) A theory of fairness, competition, and cooperation. Q J Econ 114(3):817–868
Fehr E, Fischbacher U, Gächter S (2002) Strong reciprocity, human cooperation, and the enforcement of social norms. Hum Nat 13(1):1–25. https://doi.org/10.1007/s12110-002-1012-7

Fonseca MA, Peters K (2021) Is it costly to deceive? People are adept at detecting gossipers' lies but may not reward honesty. Philos Trans R Soc Lond Ser B Biol Sci 376(1838):20200304. https://doi.org/10.1098/rstb.2020.0304
Frank SA (1998) Foundations of social evolution, vol 2. Princeton University Press. https://doi.org/10.2307/j.ctvs32rv2
Frank RH (1988) Passions within reason: the strategic role of the emotions. W W Norton & Co, pp xiii, 304
Friedman JW (1971) A non-cooperative equilibrium for supergames12. Rev Econ Stud 38(1):1–12. https://doi.org/10.2307/2296617
Fudenberg D, Rand DG, Dreber A (2012) Slow to anger and fast to forgive: cooperation in an uncertain world. Am Econ Rev 102(2):720–749. https://doi.org/10.1257/aer.102.2.720
Gavrilets S, Vose A (2006) The dynamics of Machiavellian intelligence. Proc Natl Acad Sci USA 103(45):16823–16828. https://doi.org/10.1073/pnas.0601428103
Giardini F, Balliet D, Power EA, Számadó S, Takács K (2022) Four puzzles of reputation-based cooperation: content, process, honesty, and structure. Hum Nat (hawthorne, n.y.) 33(1):43–61. https://doi.org/10.1007/s12110-021-09419-3
Gintis H (2000) Strong reciprocity and human sociality. J Theor Biol 206(2):169–179. https://doi.org/10.1006/jtbi.2000.2111
Gintis H, Bowles S, Boyd R, Fehr E (eds) (2005) Moral sentiments and material interests: the foundations of cooperation in economic life. MIT Press
Goodman JR, Ewald PW (2021) The evolution of barriers to exploitation: sometimes the red queen can take a break. Evol Appl 14(9):2179–2188. https://doi.org/10.1111/eva.13280
Goodman JR, Caines A, Foley RA (2023) Shibboleth: An agent-based model of signalling mimicry. PLoS ONE 18(7):e0289333. https://doi.org/10.1371/journal.pone.0289333
Goodman J, Crema E, Nolan F, Cohen E, Foley R (2021) Accents as honest signals of in-group membership. https://doi.org/10.33774/coe-2021-c9zhv
Guala F (2012) Reciprocity: weak or strong? What punishment experiments do (and do not) demonstrate. Behav Brain Sci 35(1):1–15. https://doi.org/10.1017/S0140525X11000069
Hagen EH, Hammerstein P (2006) Game theory and human evolution: a critique of some recent interpretations of experimental games. Theor Popul Biol 69(3):339–348. https://doi.org/10.1016/j.tpb.2005.09.005
Hamilton WD (1964) The genetical evolution of social behaviour. I. J Theor Biol 7(1):1–16. https://doi.org/10.1016/0022-5193(64)90038-4
Han TA, Perret C, Powers ST (2021) When to (or not to) trust intelligent machines: insights from an evolutionary game theory analysis of trust in repeated games. Cogn Syst Res 68:111–124. https://doi.org/10.1016/j.cogsys.2021.02.003
Hare B (2017) Survival of the friendliest: homo sapiens evolved via selection for prosociality. Annu Rev Psychol 68:155–186. https://doi.org/10.1146/annurev-psych-010416-044201
Heckathorn DD (1989) Collective action and the second-order free-rider problem. Ration Soc 1(1):78–100. https://doi.org/10.1177/1043463189001001006
Henrich J (2004) Cultural group selection, coevolutionary processes and large-scale cooperation. J Econ Behav Organ 53(1):3–35. https://doi.org/10.1016/S0167-2681(03)00094-5
Henrich N, Henrich J (2007) Why humans cooperate: a cultural and evolutionary explanation. Oxford University Press
Henrich J, Muthukrishna M (2021) The origins and psychology of human cooperation. Annu Rev Psychol 72(1):207–240. https://doi.org/10.1146/annurev-psych-081920-042106
Henrich J, Boyd R, Bowles S, Camerer C, Fehr E, Gintis H, McElreath R, Alvard M, Barr A, Ensminger J, Henrich NS, Hill K, Gil-White F, Gurven M, Marlowe FW, Patton JQ, Tracer D (2005) "Economic man" in cross-cultural perspective: behavioral experiments in 15 small-scale societies. Behav Brain Sci 28(6):795–815. https://doi.org/10.1017/S0140525X05000142
Henrich J (2020) The WEIRDest people in the world. https://us.macmillan.com/books/9780374710453/theweirdestpeopleintheworld
Hilbe C, Nowak MA, Sigmund K (2013) Evolution of extortion in iterated Prisoner's dilemma games. Proc Natl Acad Sci USA 110(17):6913–6918. https://doi.org/10.1073/pnas.1214834110
Humphrey NK (1976) The social function of intellect. In: Growing points in ethology. Cambridge University Press
Iannone R (2022) DiagrammeR. https://github.com/rich-iannone/DiagrammeR

Ibrahim AM (2022) The conditional defector strategies can violate the most crucial supporting mechanisms of cooperation. Sci Rep 12:15157. https://doi.org/10.1038/s41598-022-18797-2
Kurzban R, Houser D (2005) Experiments investigating cooperative types in humans: a complement to evolutionary theory and simulations. Proc Natl Acad Sci USA 102(5):1803–1807. https://doi.org/10.1073/pnas.0408759102
Leimar O, Hammerstein P (2001) Evolution of cooperation through indirect reciprocity. Proc R Soc Lond Ser B Biol Sci 268(1468):745–753. https://doi.org/10.1098/rspb.2000.1573
Lindenfors P, Wartel A, Lind J (2021) "Dunbar's number" deconstructed. Biol Lett 17(5):20210158. https://doi.org/10.1098/rsbl.2021.0158
Mathew S (2017) How the second-order free rider problem is solved in a small-scale society. Am Econ Rev 107(5):578–581. https://doi.org/10.1257/aer.p20171090
Mazar N, Amir O, Ariely D (2008) The dishonesty of honest people: a theory of self-concept maintenance. J Mark Res 45(6):633–644. https://doi.org/10.1509/jmkr.45.6.633
McElreath R, Boyd R, Richerson PJ (2003a) Shared norms and the evolution of ethnic markers. Curr Anthropol 44(1):122–130. https://doi.org/10.1086/345689
McElreath R, Clutton-Brock TH, Fehr E, Fessler DMT, Hagen EH, Hammerstein P, Kosfeld M, Milinski M, Silk JB, Tooby J, Wilson MI (2003) Group report: the role of cognition and emotion in cooperation. In: Genetic and cultural evolution of cooperation. MIT Press, pp 125–152
McNally L, Brown SP, Jackson AL (2012) Cooperation and the evolution of intelligence. Proc R Soc B Biol Sci 279(1740):3027–3034. https://doi.org/10.1098/rspb.2012.0206
Milinski M, Semmann D, Krambeck H-J (2002) Reputation helps solve the "tragedy of the commons." Nature 415(6870):424–426. https://doi.org/10.1038/415424a
Mokros A, Menner B, Eisenbarth H, Alpers GW, Lange KW, Osterheider M (2008) Diminished cooperativeness of psychopaths in a prisoner's dilemma game yields higher rewards. J Abnorm Psychol 117(2):406–413. https://doi.org/10.1037/0021-843X.117.2.406
Moreau L (2020) Social inequality before farming? Multidisciplinary approaches to the study of social organization in prehistoric and ethnographic hunter-gatherer-fisher societies. McDonald Institute for Archaeological Research. https://doi.org/10.17863/CAM.60627
Moya C, Boyd R (2016) The evolution and development of inferential reasoning about ethnic markers: comparisons between Urban United States and rural highland Peru. Curr Anthropol 57(S13):S131–S144. https://doi.org/10.1086/685939
Mulder LB, van Dikj E, De Cremer D, Wilke HAM (2006) Undermining trust and cooperation: the paradox of sanctioning systems in social dilemmas. J Exp Soc Psychol. https://doi.org/10.1016/j.jesp.2005.03.002
Nesse RM (2016) Social selection is a powerful explanation for prosociality. Behav Brain Sci 39:e47. https://doi.org/10.1017/S0140525X15000308
Nettle D, Dunbar RIM (1997) Social markers and the evolution of reciprocal exchange. Curr Anthropol 38(1):93–99
Noë R, Hammerstein P (1995) Biological markets. Trends Ecol Evol 10(8):336–339. https://doi.org/10.1016/S0169-5347(00)89123-5
Nowak MA, Sigmund K (1998) Evolution of indirect reciprocity by image scoring. Nature 393(6685):573–577. https://doi.org/10.1038/31225
Nowak MA, Sigmund K (2005) Evolution of indirect reciprocity. Nature 437(7063):1291–1298. https://doi.org/10.1038/nature04131
Panchanathan K, Boyd R (2004) Indirect reciprocity can stabilize cooperation without the second-order free rider problem. Nature 432(7016):499–502. https://doi.org/10.1038/nature02978
Press WH, Dyson FJ (2012) Iterated Prisoner's Dilemma contains strategies that dominate any evolutionary opponent. Proc Natl Acad Sci 109(26):10409–10413. https://doi.org/10.1073/pnas.1206569109
Raihani N (2021) The social instinct: how cooperation shaped the world. Jonathan Cape
Rand DG, Greene JD, Nowak MA (2012) Spontaneous giving and calculated greed. Nature 489(7416):427–430. https://doi.org/10.1038/nature11467
Rawls J (1955) Two concepts of rules. Philos Rev 64(1):3–32. https://doi.org/10.2307/2182230
Richerson PJ, Boyd R (2005) Not by genes alone: how culture transformed human evolution. University of Chicago Press, pp ix, 332
Riolo RL, Cohen MD, Axelrod R (2001) Evolution of cooperation without reciprocity. Nature 414(6862):441–443. https://doi.org/10.1038/35106555
Roberts G (2005) Cooperation through interdependence. Anim Behav 70(4):901–908. https://doi.org/10.1016/j.anbehav.2005.02.006

Robson AJ (1990) Efficiency in evolutionary games: darwin, Nash and the secret handshake. J Theor Biol 144(3):379–396. https://doi.org/10.1016/S0022-5193(05)80082-7
RStudio Team (2022) RStudio: integrated development for R. RStudio, PBC, Boston. http://www.rstudio.com/
Shultz S, Dunbar RIM (2007) The evolution of the social brain: anthropoid primates contrast with other vertebrates. Proc Biol Sci 274(1624):2429–2436. https://doi.org/10.1098/rspb.2007.0693
Singh M, Hoffman M (2021) Commitment and impersonation: a reputation-based theory of principled behavior. https://doi.org/10.31234/osf.io/ua57r
Singh M, Garfield ZH (2022) Evidence for third-party mediation but not punishment in Mentawai justice. Nat Hum Behav 6(7):930–940. https://doi.org/10.1038/s41562-022-01341-7
Smaldino PE, Flamson TJ, McElreath R (2018) The evolution of covert signaling. Sci Rep 8(1):4905. https://doi.org/10.1038/s41598-018-22926-1
Sperber D, Baumard N (2012) Moral reputation: an evolutionary and cognitive perspective. Mind Lang 27(5):495–518. https://doi.org/10.1111/mila.12000
Sterelny K (2021) The Pleistocene social contract: culture and cooperation in human evolution. Oxford University Press
Stewart AJ, Plotkin JB (2013) From extortion to generosity, evolution in the Iterated Prisoner's Dilemma. Proc Natl Acad Sci USA 110(38):15348–15353. https://doi.org/10.1073/pnas.1306246110
Számadó S, Balliet D, Giardini F, Power EA, Takács K (2021) The language of cooperation: reputation and honest signalling. Philos Trans R Soc B Biol Sci 376(1838):20200286. https://doi.org/10.1098/rstb.2020.0286
Tate DA (1979) Preliminary data on dialect in speech disguise. Cilt.9.90tat; John Benjamins Publishing Company. Retrieved 18 Aug 2022. https://benjamins.com/catalog/cilt.9.90tat
Traulsen A, Nowak MA (2007) Chromodynamics of cooperation in finite populations. PLoS ONE 2(3):e270. https://doi.org/10.1371/journal.pone.0000270
Trivers RL (1971) The evolution of reciprocal altruism. Q Rev Biol 46(1):25. https://doi.org/10.1086/406755
Trivers R (2000) The elements of a scientific theory of self-deception. Ann N Y Acad Sci 907:114–131. https://doi.org/10.1111/j.1749-6632.2000.tb06619.x
Tullock G (2004) Social dilemma: v. 8: of autocracy, revolution, Coup D'Etat and War: The Social Dilemma, of Autocracy, Revolution, Coup D'Etat and War v. 8 (Selected ... Of Autocracy, Revolution, Coup d'Etat & War (Volume 8 ed. edition)
Verplaetse J, Vanneste S, Braeckman J (2007) You can judge a book by its cover: the sequel. A kernel of truth in predictive cheating detection. Evol Hum Behav 28(4):260–271. https://doi.org/10.1016/j.evolhumbehav.2007.04.006
West SA, Griffin AS, Gardner A (2007) Social semantics: altruism, cooperation, mutualism, strong reciprocity and group selection. J Evol Biol 20(2):415–432. https://doi.org/10.1111/j.1420-9101.2006.01258.x
West-Eberhard MJ (1983) Sexual selection, social competition, and speciation. Q Rev Biol 58(2):155–183. https://doi.org/10.1086/413215
Wickham H, Chang W, Henry L, Pedersen TL, Takahashi K, Wilke C, Woo K, Yutani H, Dunnington D, RStudio (2021) ggplot2: create elegant data visualisations using the grammar of graphics (3.3.5) [computer software]. https://CRAN.R-project.org/package=ggplot2
Wiessner P (2005) Norm enforcement among the Ju/'hoansi Bushmen: a case of strong reciprocity? Hum Nat (hawthorne, n.y.) 16(2):115–145. https://doi.org/10.1007/s12110-005-1000-9
Wiseman T, Yilankaya O (2001) Cooperation, secret handshakes, and imitation in the Prisoners' dilemma. Games Econ Behav 37(1):216–242. https://doi.org/10.1006/game.2000.0836
Wrangham RW (2021) Targeted conspiratorial killing, human self-domestication and the evolution of groupishness. Evolut Hum Sci 3:e26. https://doi.org/10.1017/ehs.2021.20
Wrangham R (2019) The goodness paradox: the strange relationship between virtue and violence in human evolution. Vintage