Cooperation Via Intimidation: An Emergent System of Mutual Threats can Maintain Social Order

Piotr Mateusz Patrzyk¹ and Martin Takáč²

¹Faculty of Business and Economics, University of Lausanne, Quartier UNIL-Dorigny, Internef 621, CH-1015 Lausanne, Switzerland

²Centre for Cognitive Science, Faculty of Mathematics, Physics and Informatics, Comenius University, Mlynská dolina 842 48 Bratislava Slovakia

Correspondence should be addressed to piotr.patrzyk@unil.ch

Journal of Artificial Societies and Social Simulation 20(4) 5, 2017

Doi: 10.18564/jasss.3336 Url: <http://jasss.soc.surrey.ac.uk/20/4/5.html>

Received: 10-08-2016

Accepted: 18-04-2017

Published: 31-10-2017

Abstract: Can human aggressiveness promote peaceful cooperation? Despite the seeming contradiction of these phenomena, our study suggests the answer is yes. We develop two agent-based models of cooperative interactions among aggressive agents threatening each other. In Model 1, we show that aggressive displays performed by dominance-seeking individuals create a system of mutual threats that effectively enforces cooperation and inhibits agents from escalating conflicts. This happens because agents observe each other fighting, which deters them from attacking each other due to aggressive reputations. In Model 2 we extend this effect to third-party interventions showing that forming alliances makes attacks more efficient and promotes the emergence of common rules determining whom to fight against. In such a state, social order is maintained by the existence of moral alliances – groups of agents willing to fight against norm violators. In summary, we argue that reputation for toughness and the aggressive predisposition of humans could have played an important role in the evolution of cooperation and moral systems.

Keywords: Cooperation, Punishment, Revenge, Conflict, Aggression, Morality

The beginning of the slaves' revolt in morality occurs when *ressentiment* itself turns creative and gives birth to values: the *ressentiment* of those beings who, denied the proper response of action, compensate for it only with imaginary revenge.
– Nietzsche (2006 [1887], p. 20)

Introduction

- 1.1 Cooperation among genetically unrelated organisms is a puzzling phenomenon because an individual has to bear a cost to perform it while it provides a benefit in a short run that is smaller than selfishly exploiting others (Clutton-Brock 2009). It is rare among rational individuals, because even if cooperation that strives for a common goal is in everyone's interest, there is no guarantee that it will be achieved unless it is also in every individual's personal interest (Olson 1971 [1965]).
- 1.2 In this study, we focus on the role of aggressive displays in the emergence of social cooperation. Previous research has shown that humans invest their own resources to harm other individuals without any conscious self-interested motive (Fehr & Gächter 2002). We are interested in the question how such sentiments evolved and whether they foster cooperative order.
- 1.3 While there is agreement on what are the *proximate* mechanisms controlling cooperation and punishment, their *ultimate* function remains debated (Alexander 1993; Burnham & Johnson 2005; Hagen & Hammerstein 2006; Krasnow et al. 2016; Raihani & Bshary 2015; West et al. 2007, 2011). In other words, while mechanistic

explanations of (non-)cooperative decisions seem uncontroversial, the question why and how these mechanisms evolved remains unanswered. Agent-based models are well suited to contribute to such an investigation because they can illustrate how given strategies perform in different environments and what macro-level patterns emerge from local interactions (Bandini et al. 2009).

- 1.4 In general, agent-based models allow rigorous specification of and direct control over agent-internal mechanisms hypothesized to cause a certain behavioral pattern. They also allow large-scale simulations (both in terms of population size and number of interactions) unfeasible in empirical experiments. They can model behavioral patterns emerging within a fixed population (e.g., Shoham & Tennenholtz 1997) or be combined with evolutionary replication dynamics to model multi-generational populations with combinations of genetic and cultural transmission (e.g., Gilbert et al. 2006). Last but not least, the models allow simulating interventions that would be unethical in real experiments.
- 1.5 The main purpose of our research is to explore adaptiveness of different aggression strategies and their impact on rate of cooperativeness in the population. We are also interested in whether effectively spread information about an aggressive reputation can lead to systems of mutual threats and dominance hierarchies that would keep the number of actual fights low. In the next section, we review some interesting findings from experimental studies on cooperation and punishment. We distinguish two types of aggressive behavior: (1) individual aggression among partners directly involved in cooperative interactions, and (2) moralized aggression performed by third parties observing interactions of others. We argue that the uses of these behaviors constitute manifestations of dominance-seeking behavior of increasing complexity, and we investigate emergent properties of societies built on them.
- 1.6 We then formalize verbal theories proposed to explain these findings into two evolutionary agent-based models with the goal of evaluating their feasibility. Both of them model social interactions based on the use of aggressive behavior:
 - In the first model, we investigate the effectiveness of different strategies regulating whom agents attack upon encounter. Agents are paired and play the Prisoner's Dilemma (PD) Game. Depending on their strategy, they can decide to fight against their partners afterwards. We introduce 4 types of agents: those who are never aggressive (*pacifists*), those who are willing to attack partners who defected against them (*avengers*), those who are willing to attack those who cooperated with them (*bullies*) and those who always want to aggress against their partner regardless of his decision (*harassers*). When agents are observed defeating their opponents, they gain a reputation for toughness which discourages others from attacking them in the future.
 - The second model is an extension of the first one, where we investigate how patterns of interactions change when neutral observers become engaged in conflicts. Therein we introduce (morally sensitive) agents who care about the interactions of others and engage in fights after observing an interaction. We consider two additional strategies: agents who fight against defectors (*rangers*) and agents who fight against cooperators (*thugs*) also when they merely observe an interaction. That leads to group fighting and analogically to the first model, agents participating in the victorious group gain a reputation for being tough.
- 1.7 We then analyze emergent patterns of interaction within generations and also observe success of different strategies in the long evolutionary run. Based on obtained results, we argue that agents with aggressive and exploitative predispositions in their decision-making faculties can give rise to a system where violence is suppressed due to the existence of credible mutual threats. We conclude that the evolution of moral systems may in part be based on adaptations to secure agents' own interests. The agent-based models were developed using the NetLogo software (version 5.2) and all data were analyzed using R software (version 3.2.2).

The Logic of Aggression

Individual aggression

- 2.1 The supposed evolutionary reason for the existence of mechanisms motivating aggressive behavior is the need for negotiating status and forcing desirable treatment from others – in a situation of conflict of interests individuals need to fight for their position and the feeling of anger is a proximate mechanism that motivates them to achieve that (Sell et al. 2009).

- 2.2** It was observed that such aggressive behaviors are commonly used by dominant individuals in order to enforce cooperation; this stabilizes social functioning through the system of threats (Clutton-Brock & Parker 1995). However, when analyzing patterns of aggression, it is crucial to emphasize that there are no *a priori* rules regarding who is an acceptable target of violence. Experimental results show an interesting variability in how humans choose targets of aggression. It has been shown that in the context of *Public Goods Games*, some people are willing to attack those who do not contribute to the common pool, whereas others are inclined to attack prosocial contributors (Herrmann et al. 2008). While punitive attitudes are usually directed at violators of prosocial norms, it has also been observed that people spitefully attack others only in order to ensure their well-being is relatively worse (Abbink & Herrmann 2011; Abbink & Sadrieh 2009; Houser & Xiao 2010; Sheskin et al. 2014). The relative prevalence of these behaviors is supposed to be related to local competitiveness (Barker & Barclay 2016; Lahti & Weinstein 2005; Sylwester et al. 2013; see also Charness et al. 2013). Due to this variability, it has been argued that researchers need to investigate all possibilities of the direction in which aggression is inflicted (Dreber & Rand 2012).
- 2.3** A relatively common type of aggressive behavior is revenge, targeted at individuals believed to have caused some harm to the aggressor in the past. It seems to be the most hard-wired and inflexible mechanism in determining decisions to inflict harm on others, since it is automatic and can be carried out by an individual without any regard to other information (Carlsmith & Darley 2008; Carlsmith et al. 2002; Crockett et al. 2014).
- 2.4** Three main functions of revenge can be distinguished: (1) reducing the relative fitness advantage of the target, (2) changing the target's expectations about the aggressor who signals his unwillingness to accept similar treatment in the future, (3) changing audience expectations about the aggressor's status (Giardini & Conte 2015; Griskevicius et al. 2009; Kurzban & DeScioli 2013; McCullough et al. 2013; Nakao & Machery 2012; Price et al. 2002). Note that these functions do not have to be consciously realized – often the only pursued goal of an individual inflicting such a harm is to make the target suffer (Giardini et al. 2010; Andrighetto et al. 2013; Giardini & Conte 2015; see also Marlowe et al. 2011) without understanding of complex social interdependencies by avengers, their victims, and observers. All that matters is the showcase of aggressiveness through which agents prove their toughness.
- 2.5** In such a way, harming another individual serves a *self-presentational* function, that is, its purpose is to restore challenged self-esteem of the aggressor (Miller 2001). By achieving this, deterrence emerges as another function of such behavior (McCullough et al. 2013). Its adaptiveness depends on a correlation between inflicting harm and benefits received in subsequent encounters (Gardner & West 2004; Jensen 2010; Johnstone & Bshary 2004) and it will emerge only when deterrence is effective and if future interactions with the same agent are likely (Krasnow et al. 2012).
- 2.6** Because of these adaptations, willingness to harm others and fear of others being hostile is an important motivator of human action (Abbink & de Haan 2014; see also Schelling 1960). Aggression serves thereby a self-protective function – it has been shown that reputation for performing aggressive acts decreases the likelihood of being attacked by others in subsequent interactions (Benard 2013, 2015).
- 2.7** It seems that aggression is the default behavioral predisposition of humans that manifests itself in the absence of complex social organizations, as famously observed by Hobbes (2011 [1651]).
- 2.8** In Model 1, we investigate which of the aggression strategies becomes prevalent in the population and what social order emerges from interactions between aggressive agents who instigate fights after social exchanges they are involved in.

Moralized aggression

- 2.9** What is perplexing about the logic of human intervention is that people are willing to pay a cost to harm others in cases when they are not directly affected by the interaction; this is known as a third-party punishment (Fehr & Fischbacher 2004). People usually perform such behavior with a conviction that they uphold some moral norms. What benefit can humans get from becoming involved in conflicts of others?
- 2.10** Becoming involved in conflicts first and foremost provides an opportunity to signal a dominant social position by performing an aggressive act (Krasnow et al. 2016). It can also appear in case of helping one's group member or ally, or in the reciprocity case when the third parties can expect a repayment of their investment in the future (Asao & Buss 2016; Marczyk 2015; Rabellino et al. 2016). Additional benefit can be reaped by increasing one's own trustworthiness in the eyes of norm-followers. In line with this understanding, it was shown that engaging in third-party punishment increases the observer-rated likeability of an individual (Gordon et al. 2014; Jordan et al. 2016; but see Przepiorka & Liebe 2016).

- 2.11** This type of aggression is closely related to moralization of actions – targets of aggression are those who are believed to break some social or moral norms. It is typically accompanied by an aggressive use of moral language (accusing, enforcing behavior by claiming one should do it) which emerges if using it is effective in manipulating others (Cronk 1994). In that perspective, moralization should be understood as a process of attributing objective moral properties to actions in order to manipulate others with implicit threats of aggression – if one is able to present some behavior not as serving a particular interest but as serving moral principles, then the manipulation is successful (Maze 1973). An individual benefits from this by engaging third-parties in fight for his case (Petersen et al. 2013).
- 2.12** In conjunction with *coalitional psychology*, this mechanism paves the way for the emergence of large-scale moral systems by enabling humans to form and assess strength of alliances (Tooby & Cosmides 2010). In such a setting, there is a pressure for developing a common set of rules allowing for dynamic coordination among observers of conflicts (DeScioli & Kurzban 2009, 2013; DeScioli 2016).
- 2.13** Moving from a system based on individual conflicts to a system based on moralized actions evoking group conflicts requires the existence of agents that are more cognitively complex. While in the case of simple revenge it was sufficient to detect harm done to oneself and the agent who caused it, third-party punishment requires some understanding of social relations. Specifically, observers need to negotiate the status of an action – something that is absent in a world ruled by personal revenge only. This created an evolutionary pressure for developing common code and might have aided the evolution of language (Boyd & Mathew 2015).
- 2.14** The evolutionary emergence of such moral alliances is not well understood and has not been investigated in simulation work before. In order to fill this gap, in Model 2 we investigate third-party involvement in conflicts, which presents an attempt to formally examine circumstances under which these strategies can be adaptive and moral order can emerge in the community.

Model 1: Individual Aggression

Model description

- 3.1** The first model (Figure 1) investigates the role of aggressive displays in enforcing cooperation in a multi-generation evolutionary setting. Agents in each generation engage in interactions, where they can lose or gain fitness. In each interaction round, one fifth of the agents are paired randomly for interaction, with the rest of the agents evenly distributed to be observers of the interactions (8 observers assigned for each interacting pair). The interacting agents first play the Prisoner's Dilemma Game and afterwards they can decide to fight their opponent based on their fighting strategy (see below) and perceived relative strength. The fitness of the interacting agents is influenced by both the outcome of the game and a potential subsequent fight (see Appendix A for a detailed algorithm).
- 3.2** We are interested in the question of what fighting strategies are most adaptive from an evolutionary point of view. These are a hereditary trait of agents: (1) *pacifists* are agents who never fight after an interaction, (2) *avengers* fight against defectors, (3) *bullies* fight against cooperators, and (4) *harassers* fight against everyone. In the first generation, the population is composed of pacifists only. See Table 1 for comparison of agent strategies.

	Fight against			
	Cooperators	Defectors	Observed cooperators	Observed defectors
Pacifists	-	-	-	-
Avengers	-	+	-	-
Bullies	+	-	-	-
Harassers	+	+	-	-
Rangers	-	+	-	+
Thugs	+	-	+	-

Table 1: Summary of strategies of agents used in the models. They differ in whom the agents want to attack after playing the PD Game. *Rangers* and *Thugs* are present only in Model 2.

- 3.3** The result of a fight depends on relative strengths of the agents involved¹. The existence of power asymmetry among agents is the crucial element of the model: stronger agents are more aggressive, decide to fight more frequently, and are more likely to win if a fight takes place.

- 3.4** A behavioral rule for decision whether to cooperate or defect in the PD game is also related to strength, but this time to *perceived* relative strength of one's opponent: the agents cooperate when they feel weaker than their opponent and defect when they feel stronger. Thus, decisions to cooperate and decisions *not* to fight are closely connected in our model, as both of them are manifestations of weak sense of entitlement, a characteristic of people who are physically feeble (Sell et al. 2009, 2016). This belief is established probabilistically according to strength differences among contestants – see Appendix A for details.
- 3.5** The choice of such a decision rule is deliberate since we want to exclude any other factors contributing to these decisions and focus on aggressive displays only. Each time an agent loses a fight against an opponent, the loser and all observers of this fight will add the winner to their *fear_list* – a privately maintained list of 'tough' agents². When assessing the strength of an opponent, the *fear_list* biases perception, i.e. the opponents in the *fear_list* are perceived to be stronger than they are (again, for details see Appendix A). This assumption is motivated by the observation that the formation of animal dominance hierarchies depends on organisms observing each other what leads to development of expectations about their status and strength (Chase et al. 2002). In our model this emergent effect coming from observation and experience is conceptualized as deterrence – overestimation of strength of successful agents.
- 3.6** To investigate the influence of information spreading in the population besides the direct observers, the information from the *fear_list* is further passed on (gossiped) to other randomly chosen agents. After 200 interaction rounds, all the agents die and are replaced with a new generation according to the replicator equation (see Equation 2 in Appendix A), i.e. fighting strategies are spread proportionally to the fitness of their bearers in the previous generation.

Figure 1: Stages of interaction in Model 1. A population of agents is randomly divided into players and observers. Each pair has 8 observers. After the PD Game, the agents can decide to fight with each other or retreat from conflict. Observers record outcomes of fights, which allow them to estimate toughness of others.

- 3.7** The strength parameter that the agents possess can be thought of as the upper-body muscle mass, a feature argued to be the most important for estimating fighting ability (Sell et al. 2012) and it is normally distributed within population. It has been shown that the fighting ability influences beliefs of individuals about their entitlements – stronger individuals are more likely to bargain for outcomes advantageous to them and have lower thresholds for aggression (Petersen et al. 2013; Sell et al. 2016).
- 3.8** Although strength is not the only feature determining the assessment of fighting ability (see e.g., Arnott & Elwood 2009; Pietraszewski & Shaw 2015), for the sake of simplicity we assume the agents make decisions based on this value only. We further assume that each agent knows accurately their own fighting ability (but see Fawcett & Johnstone 2010; Garcia et al. 2014) and can observe the fighting ability of others, although these observations are systematically biased, as already noted (Fawcett & Mowles 2013; Prenter et al. 2006). We also assume that both contestants are equally motivated to compete for resources.
- 3.9** In this way, our model does not suffer from unrealistic assumptions found in previous models – that everybody is, in principle, able to punish everybody else³. In real life, it is not true, because individuals differ in their fighting ability. At the same time, the outcome of a fight is never certain (see Gordon et al. 2014). Agents make assessments that are probabilistic in their nature and mistakes are very common in real life (see Figures 10, 11 in the Appendix A). Perfectly rational and omniscient individuals always decide to fight when they are stronger and retreat from the conflict when they are weaker. In a population of all infallible individuals, fights would never occur. The fact they do occur suggests that individuals can make errors in their assessments, some of them resulting from adaptive biases (Johnson & Fowler 2011).

3.10 Our primary research question is which fighting strategies become dominant and which conditions promote cooperation in the PD game and reduce number of fights. We were especially interested in the impact of population size (we ran simulations for sizes equal to: 50, 100, or 500), the impact of deterrence effect strength (conceptualized as the number of standard deviations added to the opponent's strength in case he was observed to have been aggressive in the past), the impact of differing cost of punishing, and the importance of gossiping (to how many agents information about conflict outcome is passed). The number of different agents one can remember (size limit on the *fear_list*) was equal to 50 and held constant across all simulations. The independent parameters are listed in Table 2 and used in the algorithm described in Appendix A. Particular values were chosen to sample from ranges that make sense relative to one other⁴.

Parameter	Used values	Description
population	{50, 100, 500}	regulates the size of the population
punishment	9	regulates the magnitude of a punishment inflicted on the agent losing a fight
pun_cost	{3, 6, 9}	regulates the cost of inflicting punishment borne by the winner of a fight
submit_cost	6	regulates the cost of retreating from conflict by submitting to one's opponent
deterrence	{1, 3, 6}	regulates the magnitude of overestimation of the opponent's strength if he is known to be aggressive
sensitivity	0.8	regulates how sensitive the agents are to differences in strength
gossip	{5, 20}	regulates to how many individuals one can pass information about the outcome of the conflict
memory	50	regulates the number of other agents one can remember (stored in <i>fear_list</i>)

Table 2: Parameters used in the model and their values chosen for simulation runs. That gave $3 * 3 * 3 * 2 = 54$ runs in total. All simulations were run for 2000 generations.

3.11 The dependent variable was the rate of *Cooperate-Cooperate* (CC) outcomes. We put forward the following hypotheses:

- H1: *Harassers* will be an evolutionary stable strategy under all parameter combinations.
- H2: Cooperation rates will be higher in smaller groups than in bigger groups.
- H3: Cooperation rates will be higher under strong deterrence effect.
- H4: Cooperation rates will be higher when gossiping is more frequent.
- H5: Cooperation rates will be higher when the cost of punishment is low.

Results

3.12 The results of Model 1 clearly show that the most adaptive strategy across all various environmental conditions is to fight against everyone (*harasser* strategy, see Table 6 in Appendix B). This is the case because these agents establish a status of being tough most effectively due to a deterrence mechanism present in our model; this causes more of their opponents to cooperate with them. They have a particular advantage over avenger agents because they attack agents who cooperated with them, effectively showing their aggressiveness, since cooperators are mostly those who are weaker and are easy to damage (this relationship follows directly from the decision rule agents possess).

3.13 In our model it is the aggression inflicted upon weaker agents that drives the success of *harasser* agents – a very cost-effective strategy of increasing one's own status in a group, as every bully understands. *Avenger* agents lose this opportunity and cannot establish their status. Thus, we find a support for H1. Nevertheless, an unexpected finding needs to be noted – when the group size is large ($n = 500$), evolution leads to coexistence of *harasser* and *bully* agents (see Figure 3). *Bully* agents perform well too since they also fight against cooperators after defecting them, but they forego opportunities to fight after *Defect-Defect* (DD) games. It seems that the extreme aggressiveness expressed by *harasser* agents pays off in small groups where valuable information about their toughness spreads effectively. In larger groups there exists a significant group of *bully* agents because *harassers'*

Figure 2: Model 1 run with a small population size ($n = 50$) and a strong deterrence effect. Cost of punishment (pun_cost) was equal to 6, information spread (gossiping) is also high ($\text{gossip} = 20$). In such favorable conditions, aggressive agents have relative advantage over those who perform less aggression. Due to the existence of mutual threats, a relatively high rate of CC outcomes (75.06 % during last ten interactions) is promoted.

costly fighting after DD games where both contestants are of similar strength is not so strongly compensated by gaining a tough reputation as it is in smaller populations.

Hypothesis	Result	Comment
H1: <i>Harassers</i> will be an evolutionary stable strategy under all parameter combinations.	Confirmed	The <i>harasser</i> strategy is most adaptive since these agents engage in fights most frequently and build the reputation for being tough
H2: Cooperation rates will be higher in smaller groups than in bigger groups.	Confirmed	Cooperation is stronger in smaller groups because agents observe each other fighting frequently what leads them to think others are tough; this is not the case in big groups with anonymous interactions
H3: Cooperation rates will be higher under strong deterrence effect.	Partially confirmed	Deterrence promotes cooperation in small groups where agents observe each other frequently; in large groups where repeated interactions are not likely its effect becomes negligible (see also Figure 4)
H4: Cooperation rates will be higher when gossiping is more frequent.	Not Confirmed	This parameter did not significantly affect results of our simulations
H5: Cooperation rates will be higher when the cost of punishment is low.	Not Confirmed	This parameter did not significantly affect results of our simulations

Table 3: Hypotheses tested in Model 1.

3.14 Figures 2-4 illustrate simulation dynamics in two different runs. As can be seen in Figure 2, in small groups the CC rate increases as the proportion of aggressive individuals increases. However, this is not the case for large groups (Figure 3). This difference is due to the availability of information about other agents in small, but not large, populations.

3.15 In large populations the lack of information about other agents' past behavior in most encounters leads to a strict dominance hierarchy – most interactions are CD outcomes (Figure 3). This happens because agents assess each other relatively accurately and “correctly” play Defect (the stronger player) and Cooperate (the weaker player). Such interactions are often followed by the weaker agent submitting (Dominations, Figure 4). Interestingly, due to dominance hierarchies, the rate of post-game fights remain low.

3.16 In smaller populations agents have information about others' past behavior in most cases and adjust their decisions accordingly. Because more of the assigned partners are remembered to be aggressive, agents overestimate their fighting ability and “incorrectly” play Cooperate. When both players fear their opponent, it leads to a CC outcome. This situation corresponds to the Hobbesian state of war in which agents live in a system of

Figure 3: Model 1 run with the large population size ($n = 500$) and strong deterrence effect. The remaining parameters are the same as in the run depicted in Figure 2. As can be seen, under conditions of anonymity, systems of mutual threats cannot be established, and in most interactions the stronger agent exploits the weaker (CD outcome rate is equal to 58.87% during last ten interactions).

Figure 4: Fighting frequency in two simulation runs. The plot on the top depicts frequencies of conflict resolution in a run with a small population size (see Figure 2), the plot on the bottom depicts a simulation run with a large population size (see Figure 3). Pacifist is the only non-aggressive strategy – as it declines, aggressive agents (avengers, bullies, and harassers) invade the population and dominance hierarchies are established.

mutual threats. The rate of fighting is similarly low as in large populations, the rate of dominations is also lower (because agents fear each other, conflicts are instigated less frequently).

3.17 Figure 5 summarizes the rates of CC at the end of different simulations. A baseline to which these rates are compared is a situation when deterrence parameter is set to 0. In this case, agents do not develop reputation for being tough so the mechanism proposed by us is not at work. Under the assumption that agents can classify themselves as either weaker or stronger without making errors, this baseline would be equal to 0, as the only possible outcome would be CD. Because we made more realistic assumption about messy classifier (see Appendix A), random errors lead also to CC and DD outcomes. We verified this by running our model with deterrence parameter set to 0 and the average CC rate achieved was equal to 19.79%⁵.

3.18 As can be seen in Figure 5, our results indicate an interaction between the group size and deterrence. Effective threats can deter defections and sustain a relatively high ratio of CC outcomes in a population when its size is small. We thus find support for H2 and a limited support for H3. When the population size is large, the deterrence effect becomes irrelevant, because most interactions will be anonymous. In that case the prevailing outcome of PD Game will be *Cooperate-Defect* (CD, see Figure 3). For the effectiveness of aggressive displays, it needs to be the case that the information is passed to other agents. The group size and the effectiveness of deterrence are contributing to the maintenance of social order. However, we do not find support for hypotheses H4 and H5, as the effects of gossiping and the cost of punishment parameters were negligible (see Appendix B).

Discussion

3.19 The results show that the existence of aggressive agents does not lead to a higher number of fights. This is the case because agents who have information about each other's toughness develop dominance hierarchies

Figure 5: CC rates at the end of simulation runs. The plot shows CC rates achieved in population based on its size and the deterrence effect. Results are averaged over runs with different gossip and punishment cost values. Black dashed line shows the baseline CC rate that is achieved without the deterrence mechanism (see also Figure 10 in Appendix A).

that allow them to resolve conflicts without the need to fight. Such dominance hierarchies are frequently observed in various animal societies (Franz et al. 2015; Hsu et al. 2006) and our model illustrates their possible evolutionary emergence.

- 3.20** Our results are also consistent with findings showing that sociality promotes suppression of fights and that conflicts are resolved through submissive displays (Hick et al. 2014).
- 3.21** How is it possible that cooperation is promoted based on a decision rule that promotes defection against everyone who is thought to be weaker? The only way to achieve it is in a situation of mutual fear – when two agents meet each other and both of them think the opponent is stronger. This follows from the formula for estimating the opponent’s strength (see Appendix A). If the two agents have each other in their *fear_lists*, their perception of the other agent’s strength will be biased by the deterrence value, which causes the stronger agent to choose cooperation and decide not to fight because of its false belief. It follows then that the only way to enforce cooperation is to make oneself stronger in the eyes of others. We assume that it can be achieved through aggressive displays which deter future partners from defecting. Deterrence is conceptualized as a tendency to overestimate the strength of individuals that have been observed to be aggressive. We show that this effect is especially pronounced in small groups (see Figure 5).
- 3.22** In this model, we provide a novel way of conceptualizing deterrence in agent interactions. Previous approaches modeled deterrence in the context of the Public Goods Game by having agents develop a reputation for being a punisher, which in turn enforced contribution when agents observed these punishers in their groups (e.g., dos Santos & Wedekind 2015; Krasnow et al. 2015). In another related model, Nowak et al. (2016) investigated the adaptiveness of different strategies regulating behavior in response to being challenged to fight. A novel feature, missing in their approaches, is the linking of these predispositions to cooperative interactions which, in our model, also affect agents’ fitness. Through this linking, our model focuses on emergent mechanisms curbing agents’ exploitative and aggressive predispositions, which were assumed to be universal.

Model 2: Moralized Aggression

Model description

- 4.1** Model 2 (see Figure 6) is an extension of Model 1 in that some of the observers (so called *morally sensitive*) will not remain passive bystanders but will actively take sides and join the fight. The outcome of the fight is then determined by relative total strengths of the two fighting groups. The model investigates the adaptiveness of different strategies regulating side-taking rules in conflicts. We introduce two new types of (morally sensitive) agents: (1) *Rangers* who engage in fights as a third-party supporting cooperators, and (2) *Thugs* who support defectors.
- 4.2** The latter strategy can be understood as a mafia-like alliance in which antisocial individuals will join their forces to inflict harm and suffering even more effectively. These agents will also possess a different decision rule whether to cooperate or defect in the PD game: in order to defect, they need not be stronger themselves; it

suffices if the number of antisocial fighters in the society is high enough to support them: another feature that, this time, a group of bullies perfectly understands⁶.

- 4.3** The *ranger* strategy can be understood as a moral alliance, which demands moral integrity from its members. Thus, the decision rule for these agents is also different. We investigate three possibilities regarding how these agents can behave: (1) they cooperate with everyone (*Always Cooperate*); (2) they defect against the agents who have been observed defecting in the past (*Defect Defectors*⁷); (3) they defect against defectors and against the agents who remain neutral in conflicts (*Defect Defectors and Neutrals*), thereby creating an additional pressure against the agents who do not care about interactions they are not involved in. The algorithmic details of Model 2 are described in Appendix C (see also Table 4 for newly introduced parameters).

Figure 6: Stages of interaction in Model 2. Roles are assigned analogically to Model 1. The behavior of observers depends on their strategy – morally insensitive agents remain neutral just like in the previous model, while morally sensitive agents will choose sides in conflict; this leads to its escalation and coalitional fights.

Parameter	Used values	Description
coal_sensitivity ⁸	0.3	regulates how sensitive the agents are to differences in the coalitional strength. Since the variance of strength differences is higher in case of fighting coalitions, this number was set to be smaller than sensitivity
Rule	{alwaysC, DefectD, DefectD&N}	regulates the decision rule for rangers: 1) Always Cooperate; 2) Defect against agents who have been observed defecting in the past; 3) Defect against agents who have been observed remaining neutral at conflicts

Table 4: Additional parameters used in Model 2. Due to their low importance in Model 1 (see Appendix B), we decided not to vary *pun_cost* and *gossip* parameters. Their value was set to 6 and 20, respectively. *Population* and *deterrence* parameters were varied analogically to Model 1. That gave $3 * 3 * 3 = 27$ runs in total. All simulations were run for 2000 generations.

- 4.4** Our main research question for Model 2 is which conditions make an intervention performed by a morally sensitive third-party adaptive. To the best of our knowledge, there are no models that would explicitly investigate this effect⁹.
- 4.5** The logic behind a potential success of the intervention strategies relies on several factors. The first is that group fighting can decrease the cost borne by the participating individuals (Jaffe & Zaballa 2010; Okada & Bingham 2008). The second factor is that these strategies allow establishment of toughness more easily, especially for weaker agents who can display aggressiveness in coalitions supporting them. Third, fighting coalitions can more easily eliminate lone agents – while this can have a destructive effect on the society, it can very often increase the relative fitness of attackers. Finally, the more individuals are involved in punishment, the stronger the deterrence effect appears to be, even if material losses are the same (Villatoro et al. 2014).
- 4.6** We put forward the following hypotheses:

- H6: *Thug* will be an evolutionary stable strategy under *Always Cooperate* decision rule, or when the population size is large.
- H7: Under *Defect Defectors* and *Defect Defectors and Neutrals* decision rules, *Ranger* will be an evolutionary stable strategy, but only when the population size is small ($n = 50$ and $n = 100$).

Results

- 4.7** In accordance with hypothesis H6, results of Model 2 show that, under ranger decision rule *Always Cooperate*, the only possible outcome is that *thugs* invade the population and the rate of DD outcomes approaches 100%, leading to costly private fights among all members of the population (see Appendix D). *Rangers* maintain their integrity by having a cooperative reputation; they inflict punishment on defectors, but this is not enough to outweigh benefits acquired by the defectors.

Hypothesis	Result	Comment
H6: <i>Thug</i> will be an evolutionary stable strategy under rangers' <i>Always Cooperate</i> decision rule, or when the population size is large.	Confirmed	Unconditional cooperation makes agents an easy target of exploitation that leads to the success of <i>thugs</i> . This happens both in the case of cooperating with everyone by default (<i>Always Cooperate</i> rule) and when interactions are anonymous (as in large groups).
H7: Under <i>Defect Defectors</i> and <i>Defect Defectors and Neutrals</i> decision rules, <i>Ranger</i> will be an evolutionary stable strategy, but only when the population size is small ($n = 50$ and $n = 100$).	Partially confirmed	This can happen, but success is not always guaranteed (see Figure 8 and Sections 4.11-4.14). Successful order maintenance happens due to the existence of a moral alliance between <i>rangers</i> .

Table 5: Hypotheses tested in Model 2.

- 4.8** In the case of the other two rangers' decision rules, when the group size is large ($n = 500$), *thug* agents also invade the population regardless of the values of remaining parameters (see Figure 7). This happens because *rangers* are not able to keep all individuals in memory, and under these circumstances of anonymity they play *Cooperate* in cases when they should not, also leading to their demise. Thus, we find full support for H6.

Figure 7: Model 2 run with large population size ($n = 500$), *Defect Defectors* behavioral rule and moderate deterrence effect. *Thug* agents invade the population and can safely play *Defect* against others since they know they are in majority. This leads to costly private fights among them (97.6% of games are DD during last ten interactions) but there is no mechanism that would give advantage to other strategies.

- 4.9** In the case of smaller group sizes, results are sensitive to random differences at the beginning of each simulation run. Figure 8 depicts a simulation run where *ranger* invaded a population promoting very high rate of CC outcomes. Such a result, however, was not obtained in all parameter combinations (see Table 7 in Appendix D). In order to verify this seemingly random pattern of results, we ran Model 2 again for small and moderate group

sizes, as well as *Defect Defectors* and *Defect Defectors and Neutrals* decision rules and now each parameter combination was repeated 10 times (with different random generator seeds).

- 4.10** Figure 9 depicts the number of simulations in which these strategies were successful. There appears to be a threshold above which one of the coalitions gets strong enough to intimidate other players in all interactions and spreads very quickly. Once one of the strategies becomes prevailing, it is evolutionary stable. Thus, while it seems that a small group size and efficient information spread help *rangers* enforce cooperative order, we do not find full support for H7 – under favorable conditions, this strategy invades the population only 65% of time. Further research needs to identify different factors contributing to the success of emerging coalitions.

Figure 8: Model 2 run with medium population size ($n = 100$), *Defect Defectors* behavioral rule and moderate deterrence effect. A relatively high rate of CC outcomes is sustained (86.25% during last ten interactions) because ranger agents are in majority, thereby deterring others from playing Defect.

Figure 9: Invasions by *rangers* and *thugs*. Each combination of parameters was run 10 times. The plot shows the number of successful invasions by each strategy, as dependent on population size, behavioral rule, and deterrence effect strength (d).

Discussion

- 4.11** The results of our simulation suggest there are two factors necessary for the emergence of moral alliances: (1) small group size where agents are not anonymous, and (2) common rules allowing for *justified* defection (defecting those who have defected, see also Axelrod & Hamilton 1981). The maintenance of social order in runs where *rangers* prevailed suggests a possible mechanism in which threat of aggression serves a prosocial role. Still, even if these conditions hold, the success of prosocial moral strategies (*rangers*) is not guaranteed.
- 4.12** In this model, enforcement is based on the presence of individuals willing to fight for a given norm. The success of moral alliance composed of *rangers* shows the paradox inherent in the use of morality – that the weapon originally invented to fight against *others* by constraining *their* behavior turns against its inventors who need to constrain their own behavior as well. We believe that the decision rule possessed by those agents exemplifies the central mechanism of morality – they refrain from exploiting others not because they expect to be beaten (as did the agents present in Model 1), but because they act according to norms (e.g., defecting *only* against defectors). Adherence to this alliance enforces cooperation on other non-moral agents when they believe they are in a minority, what leads to even higher CC rates than those observed in Model 1.

- 4.13** In this model we formalized the verbal argument regarding the emergence of impartial morality put forward by DeScioli & Kurzban (2009, 2013); DeScioli (2016). We thus provide some support for their proposal by showing that agents committing themselves to a moral alliance and upholding norms agreed upon by them can perform well despite their refraining from exploitation of others. This happens because they effectively bring order into society and promote cooperation.
- 4.14** However, it is not clear what other factors contribute to the success of *rangers* as opposed to *thugs*. In our model the success of *rangers* is not guaranteed. It suggests that other factors such as social exclusion of non-cooperative individuals and forming closed communities with a limited number of partners might also be important in the emergence of cooperation (see e.g., Baumard et al. 2013; Helbing et al. 2010).

General Discussion

- 5.1** In the current study, we developed two agent-based models of social conflict, each leading to a system of social aggression with characteristic complexities. The models illustrate a way in which social order can be maintained among aggressive agents with minimal assumptions regarding their cognitive capacities.
- 5.2** In the first model, we assumed that agents seek dominance and use aggression in order to fight for resources. Their social life was limited to observing fights performed by others, which helped in making their own decisions regarding escalating conflicts. The results of this model showed that this system would lead to vengeance and social exploitation of weaker agents by those who had power to do so. Despite such a grim world, we showed that the system of mutual threats established by everyone's aggressiveness could constitute an effective deterrent discouraging the agents from seeking dominance violently. Interactions of such intimidated agents produced a relatively high level of cooperative outcomes.
- 5.3** We believe that such understanding of evolutionary roots of norm-related behavior sheds light on the existence of moral hypocrisy in social interactions (see Batson 2016). Rather than conceptualizing morality as a device for controlling one's own behavior, we conceptualized it as a device for controlling others' behavior. As we argued, this presents a more reasonable approach to the problem from the evolutionary perspective. We have shown that in a group of agents possessing primarily aggressive and other-directed morality, adherence to moral rules can nevertheless be promoted. Adherence happens not because agents intrinsically value adherence to moral rules, as there are no evolutionary reasons for such a sentiment to be adaptive. Instead, adherence happens because agents fear others who are willing to defend *their* (other-directed) moral rules.
- 5.4** Such a state makes moral hypocrisy an adaptive strategy (when unobserved, agents would try to break the rules themselves) and creates a pressure for the development of behavioral strategies that control one's own behavior depending on whether others are observing the agent (see also Trivers 1971). We think that such framing illuminates the roots of *defensive* morality which is concerned with securing a cooperative reputation within a group (Alexander 1987). An emergent property of a system based on aggression is that in the presence of individuals willing to fight for a given norm it is adaptive to follow the norm, because it diminishes the probability of being attacked (or, to be more precise, it is adaptive to *appear* to be a norm follower in the eyes of others).
- 5.5** In the second model, we made an additional assumption regarding members of the society – we endowed agents with the ability to form alliances and wage group wars. We showed that such a system established a selection pressure for developing common rules for forming alliances and conflict resolution. Our results illustrate that in the absence of moral alliances, societies are invaded by antisocial groups (*thugs*), whose interactions lead to mutual defections and costly fights. They also show that under favorable circumstances, reliable coalitions can enforce cooperative outcomes. For this to happen, it is necessary for agents promoting moral order to allow for justified defections according to widely accepted norms.
- 5.6** In such a way, our study builds on the philosophical argument on the emergence of morality developed by Nietzsche (2006 [1887]). Cooperative behavior in the Prisoner's Dilemma Game is usually thought to be a moral or socially desirable response, while decision to defect is thought to be a selfish or at least distrustful choice. However, on a more basic level, these decisions can be analyzed not on a cooperative versus non-cooperative continuum, but on an inferiority versus assertiveness continuum. It is primarily not a moral choice, but a choice between asserting one's own superiority and aiming for higher payoffs (defecting) or relinquishing potential payoff to the partner (cooperating). Cooperative behavior in our first model was not chosen by agents who were the fittest, it was chosen by those who were weak and intimidated. Such a conceptualization sheds light on findings suggesting that strictness in moral judgments about one's own transgressions is typical for individuals who have low power, as they do not feel entitled to fight for their interests (Lammers et al. 2010).

- 5.7** In Nietzsche's terms, moral prescription to cooperate has its roots in the *slave morality*. It is not based on striving to be superior; it is based on being inferior. The evolutionary emergence of moralization (the *slave's revolt*) happens at the moment when agents following cooperative strategies (*rangers*) form an alliance in which they enforce cooperative norms on others. This corresponds to the moment of transition from the system where cooperation was a necessary manifestation of inferiority when one feared his opponent (Model 1) to the system where, due to the newly-formed alliance, the concept of behavioral consistency enters the picture and *rangers* need to practice what they preach by following cooperative norm themselves (Model 2).
- 5.8** We would like to conclude our paper with emphasizing some important limitations of our results. While we found support for our hypotheses regarding evolutionary pressures created by modeled mechanism, in both models these failed to promote cooperative outcomes for larger populations. This suggests that there are likely also other mechanisms at play in promoting cooperative outcomes. Specifically, the reason why we obtained such results is that we assumed interactions in well-mixed populations, where each agent is equally likely to be paired with every other member of the group. Modeling variations on this grouping scheme was beyond scope of this study but it could be expected that taking into consideration local interactions within larger population would make information spread more efficient and promote more cooperative outcomes. These effects are not inconsistent with the framework outlined here and future work needs to focus on them in investigating the emergence of moral alliances.

Acknowledgements

We would like to thank the anonymous reviewers for their helpful comments. Figures 1 and 6 contain graphics designed by Freepik.

Appendix A: Description of Model 1

In this appendix, we provide an algorithmic description of the first model developed in the study.

```

1 generation ← 1;
2 while generation ≤ max_generation do
3   A new generation of agents is born. In the first generation, all agents are pacifists;
4   for i ← 1 to 200 do
5      $\frac{1}{5}$  of agents are chosen randomly to be players of the PD Game;
6     Players are paired and for each pair, 8 randomly chosen agents from the remaining pool are chosen to
       be observers;
7     for each pair do
8       for each player do
9         if opponent ∈ fear_list then
10          | bias ← deterrence;
11         else
12          | bias ← 0;
13          | diff ← own_strength − (partner_strength + bias);
14          The player thinks he is stronger with probability returned by the sigmoid function (1);
15          if think I am stronger then
16          | play defect;
17          else
18          | play cooperate;
19        end
20      Agents play the PD Game, the payoff is determined by the following table and added to agents'
        fitness:
        

|           |           |        |
|-----------|-----------|--------|
|           | cooperate | defect |
| cooperate | 6,6       | 0,10   |
| defect    | 10,0      | 2,2    |


21      Each agent can decide to attack his partner - this decision depends on the strategy and is
        determined by the following table (for row player):
        

|           |                        |                          |
|-----------|------------------------|--------------------------|
|           | cooperate              | defect                   |
| cooperate | -                      | -                        |
| defect    | <i>bully, harasser</i> | <i>avenger, harasser</i> |


22      if both players decide to fight then
23        The winner is established with the probability returned by the sigmoid function (true strengths
        taken as input, i.e. without bias);
24        The winner pays pun_cost;
25        The loser pays punishment;
26      else if one of them decides to fight and the other does not then
27        The non-aggressive individual retreats and pays submit_cost;
28        The loser and the observers add the winner to their fear_lists;
29        gossip agents are chosen randomly for each observer, and the winner is added to their fear_lists;
30      end
31    end
32    The current generation of agents dies and replicates according to the replicators dynamics (2) ;
33    Each newly born agent mutates with the probability  $p = 0.005$  (his strategy is changed to a different,
    randomly chosen one);
34    generation ← generation + 1;
35 end

```

$$sigmoid = \frac{1}{1 + \exp\left(-diff \cdot sensitivity\right)} \quad (1)$$

$$Pr_{t+1}(i) = \frac{Pr_t(i) \cdot F_t(i)}{\sum_{j=1}^N Pr_t(j) \cdot F_t(j)} \quad (2)$$

Equation 1 depicts the sigmoid function that returns the probability of feeling stronger. Equation 2 depicts the Replicator Dynamics used in the model. $Pr_t(i)$ denotes the proportion and $F_t(i)$ denotes the average fitness of agents following strategy i at time t .

Figure 10: The probability of deciding to fight – the plot shows the output of the sigmoid function 1 for different *sensitivity* values: 1) Random Classifier, *sensitivity* = 0; 2) Perfect Classifier, *sensitivity* = $+\infty$; 3) Classifier used in the model, *sensitivity* = 0.8.

Figure 11: Distribution of the PD Game outcome. The figure presents a theoretical distribution of outcomes under the use of different sensitivity values (see Figure 10). Rates for deterrence parameter depict a theoretical case in which everyone fears everyone.

Appendix B: Results of Model 1

Table 6 presents the full results of Model 1 runs:

popula tion	pun_cost	gossip	deter- rence	paci- fists	aven- gers	harassers	bullies	DDs	CDs	CCs
50	3	5	1	10	8	74	8	10.2	57.77	32.03
50	3	5	3	2	4	88	6	2.95	38.49	58.56
50	3	5	6	4.2	2	84.2	9.6	1.83	26.79	71.38
50	3	20	1	10	12.6	64.6	12.8	9.85	55.46	34.69
50	3	20	3	4.4	6	76	13.6	2.56	34.81	62.63
50	3	20	6	5.6	7	77.8	9.6	1.56	24.03	74.41
50	6	5	1	6	3.4	73	17.6	9.36	56.86	33.78
50	6	5	3	4	12.6	67.4	16	3.11	38.58	58.31
50	6	5	6	4.6	5.4	78	12	1.95	28	70.05
50	6	20	1	9.2	14.6	63	13.2	9.63	56.64	33.73
50	6	20	3	3	7	72.8	17.2	2.33	35.85	61.82
50	6	20	6	2	2.8	89.2	6	1.28	23.66	75.06
50	9	5	1	16	8.8	51.2	24	10.44	57.96	31.6

Continued on following page

50	9	5	3	5.8	4	74.6	15.6	3.36	38.42	58.22
50	9	5	6	3.8	4	84.2	8	2.12	27.15	70.73
50	9	20	1	11.8	14	18	56.2	9.9	57.13	32.97
50	9	20	3	3.8	7.2	77.6	11.4	2.45	35.1	62.45
50	9	20	6	0	6	87.4	6.6	1.36	23.6	75.04
100	3	5	1	7.9	9	54.7	28.4	13.71	58.335	27.955
100	3	5	3	5.2	6.3	75.9	12.6	7.435	51.465	41.1
100	3	5	6	5.6	5.2	81.9	7.3	5.955	45.8	48.245
100	3	20	1	9	7.9	65.5	17.6	13.52	58.495	27.985
100	3	20	3	3.9	5.3	84.2	6.6	7.455	51.325	41.22
100	3	20	6	2	3	90.3	4.7	5.685	45.855	48.46
100	6	5	1	10	8.4	49.2	32.4	13.59	58.18	28.23
100	6	5	3	8.1	4.7	76	11.2	7.405	51.815	40.78
100	6	5	6	4.7	4.2	77.5	13.6	6.055	45.97	47.975
100	6	20	1	10.1	6.8	65.5	17.6	13.135	59.495	27.37
100	6	20	3	4.8	5.1	79.8	10.3	7.415	50.865	41.72
100	6	20	6	3.6	4.6	82.7	9.1	5.775	45.28	48.945
100	9	5	1	9.5	11.2	42.7	36.6	13.47	58.955	27.575
100	9	5	3	2.3	4.2	80.9	12.6	7.53	51.715	40.755
100	9	5	6	3	3.5	84.7	8.8	5.895	45.155	48.95
100	9	20	1	4.7	9.2	46	40.1	13.235	59.115	27.65
100	9	20	3	3.4	6.2	73.6	16.8	7.235	50.945	41.82
100	9	20	6	3.5	4.7	79.7	12.1	6.08	45.97	47.95
500	3	5	1	13.5	17.54	47.42	21.54	18.368	60.754	20.878
500	3	5	3	8.24	10.4	63.56	17.8	16.704	60.263	23.033
500	3	5	6	6.6	7.02	70	16.38	16.634	58.965	24.401
500	3	20	1	12.76	16.96	45.04	25.24	18.506	60.219	21.275
500	3	20	3	8.04	8.58	60.34	23.04	17.065	59.551	23.384
500	3	20	6	9.8	8.36	67.72	14.12	16.45	59.134	24.416
500	6	5	1	17.38	11.78	30.16	40.68	18.366	60.251	21.383
500	6	5	3	11.12	7.44	38.66	42.78	17.034	59.735	23.231
500	6	5	6	7.4	8	52.9	31.7	16.316	59.205	24.479
500	6	20	1	17.72	10.46	27.46	44.36	18.304	60.512	21.184
500	6	20	3	12.04	8.46	42.12	37.38	16.971	59.448	23.581
500	6	20	6	7.68	9.34	50.22	32.76	16.72	58.872	24.408
500	9	5	1	21.42	13.18	20.94	44.46	18.411	60.33	21.259
500	9	5	3	9.38	9.88	30.38	50.36	17.098	59.74	23.162
500	9	5	6	8.28	7.2	38.42	46.1	16.081	59.77	24.149
500	9	20	1	19.36	9.76	24.24	46.64	18.271	60.617	21.112
500	9	20	3	12.46	9.28	30.9	47.36	17.086	59.573	23.341
500	9	20	6	9.82	6.86	35.02	48.3	16.736	58.975	24.289

Table 6: Strategy frequencies and PD Game outcomes at the end of simulation runs. The table shows how average relative frequencies of strategies and game outcomes in the last 10 generations depend on parameter values (gray column heads).

Appendix C: Description of Model 2

In this appendix, we provide a detailed description of the second model developed in the study. The structure of interactions is related to those present in Model 1. Agents are assigned to roles of players and observers just like in Model 1. The difference is in the way types of agents introduced here make decisions regarding playing the PD Game, see lines 9-18 of the algorithm in the Appendix A.

Ranger agents maintain in memory decisions made by other players by updating *defector_list* and *neutral_list* (players following strategies present in Model 1 who do not engage in group fights).

When *rule* parameter is set to *alwaysC*, *rangers* always cooperate with the partner.

When it is set to *DefectD*, rangers make the decision in a following way:

```

1 if opponent ∈ defector_list then
2   |   play defect;
3 else
4   |   play cooperate;
5 end

```

When it is set to *DefectD&N*, rangers make the decision in a following way:

```

1 if opponent ∈ defector_list OR opponent ∈ neutral_list then
2   |   play defect;
3 else
4   |   play cooperate;
5 end

```

Thug agents maintain an estimate of the frequency of their strategy relative to *rangers* (*rangers/thugs*). In the first generation, they are born with an estimate equal to 0.9, which is then subsequently updated as agents observe new contests. This value was chosen such that thugs receive a slight advantage at the beginning of simulation – they decide to defect in their first interactions and stop doing that only if they learn they are in minority (see below).

Within the generation, this estimate is updated in each interaction round.

$$new_ratio = \left(0.8 \times old_ratio\right) + \left(0.2 \times \frac{ranger}{thug}\right) \quad (3)$$

ranger denotes the number of altruistic fighters (supporting the cooperator) perceived in this round, whereas *thug* denotes the number of antisocial fighters (supporting the defector) perceived in this round.

In subsequent generations, *thug* agents are born with the estimate equal to the average estimate present in the previous generation.

Thug agents make a decision how to play the PD Game, according to the following rule:

```

1 if ratio < 1 then
2   |   play defect;
3 else
4   |   play cooperate;
5 end

```

For the punishment stage of the game (lines 21-29 of the algorithm in Appendix A), there are two possibilities: either there are no morally sensitive agents among observers, in which case the conflict resolution will be based on a private fight just like in Model 1, or there are morally sensitive observers, and that will lead to a coalitional fight.

Figure 12: The cost of group punishment depending on the number of aggressors.

In the latter case, the procedure is the following:

```

1 if agents played Cooperate-Defect then
2   for each observer do
3     if strategy = ranger then
4       support Cooperator;
5     else if strategy = thug then
6       support Defector;
7     else
8       remain neutral;
9   end
10 else if agents played Defect-Defect then
11   if one of the players is ranger while the other is not then
12     for each observer do
13       if strategy = ranger then
14         support ranger player;
15       else if strategy = thug then
16         support the other player;
17       else
18         remain neutral;
19     end
20   else
21     the conflict is resolved privately just like in the Model 1;
22   end
23 for each coalition do
24    $diff \leftarrow \left( \sum we\_strengths \right) - \left( \sum they\_strengths \right);$ 
25   The coalition members think they are stronger and are willing to fight with the probability returned by the
   sigmoid function (1) (coal_sensitivity used);
26 end
27 if one coalition is willing to fight while the other is not then
28   the agent involved in the interaction supported by retreating coalition is attacked and pays punishment;
29 else if both coalitions are willing to fight then
30   the winner is established with the probability returned by the sigmoid function (1);
31   all members of the losing coalition pay punishment;
32 cost borne by members of the winning coalition is determined by equation (4);
33 all members of the losing coalition and all neutral agents add every member of the winning coalition to their
   fear_lists and pass this information to gossip agents;
34 (rangers only) lists of defectors and neutrals are updated;
35 (thugs only) ratio  $\frac{rangers}{thugs}$  is updated;

```

$$cost = \frac{pun_cost}{\#members^2} \quad (4)$$

Equation 4 shows how the cost borne by one group member is calculated. Agents replicate and mutate in the same way that in Model 1, see Appendix A, the algorithm lines 32-33.

Note that the equation determining decisions to fight (Equation 1, Appendix A) in case of group conflicts uses different sensitivity value than in the case of individual conflicts. This is because strength differences between groups of individuals can be higher than differences between single individuals – in other words, the range of inputs to the sigmoid function increases (see Figure 10). For that reason, in order to keep the baseline of random cooperation similar to the level found in individual contests (see Figure 5), we needed to change the value sensitivity parameter. Assigning higher value would very slightly decrease number of CC outcomes, but would not qualitatively change our results.

Appendix D: Results of Model 2

Table 7 presents the full results of Model 2 runs:

rule	popula- tion	deter- rence	paci- fists	aven- gers	haras- sers	bullies	rangers	thugs	DDs	CDs	CCs
AC	50	1	2.2	0	3	0	2	92.8	89.52	10.29	0.19
DD	50	1	7	2	3.6	6.6	74.8	6	1.36	14.99	83.65
DD&N	50	1	3	2	2	2	6.2	84.8	88.4	11.14	0.46
AC	50	3	2.4	0	2.6	0.4	2	92.6	88.14	11.65	0.21
DD	50	3	6	0	6	11.4	67.2	9.4	0.64	11.41	87.95
DD&N	50	3	2.2	2	3	4	82.4	6.4	2.18	21.8	76.02
AC	50	6	0	0	0	2.2	0.4	97.4	95.55	4.45	0
DD	50	6	1.2	1.8	1.4	2	2.8	90.8	87.33	12.24	0.43
DD&N	50	6	4	0	3.4	2.8	88.6	1.2	0.88	13.8	85.32
AC	100	1	1.1	1	1.3	0	1	95.6	93.925	6.045	0.03
DD	100	1	4.9	4.4	1.6	3.1	78.1	7.9	1.355	12.21	86.435
DD&N	100	1	1.8	1.5	1	1	92.2	2.5	2.595	7.46	89.945
AC	100	3	1.3	1	1	1.2	1.2	94.3	91.595	8.31	0.095
DD	100	3	7	3	3.2	6	74.1	6.7	1.41	12.345	86.245
DD&N	100	3	1.3	1.4	1	0	1.9	94.4	94.01	5.88	0.11
AC	100	6	1.5	0.5	1	1.2	0.9	94.9	92.16	7.775	0.065
DD	100	6	1.2	1	1	1.1	1	94.7	93.08	6.83	0.09
DD&N	100	6	1.2	2.7	2.5	2.3	87.1	4.2	2.745	14.06	83.195
AC	500	1	0.56	0.6	0.78	0.36	0.26	97.44	96.992	2.996	0.012
DD	500	1	0.4	0.6	0.54	0.74	0.24	97.48	97.137	2.844	0.019
DD&N	500	1	0.64	0.4	0.5	0.36	0.3	97.8	97.604	2.387	0.009
AC	500	3	0.58	0.5	0.28	0.4	0.34	97.9	97.578	2.407	0.015
DD	500	3	0.36	0.48	0.54	0.38	0.26	97.98	97.6	2.389	0.011
DD&N	500	3	0.52	0.84	0.36	0.34	0.3	97.64	97.349	2.629	0.022
AC	500	6	0.7	0.52	0.56	0.46	0.36	97.4	96.833	3.144	0.023
DD	500	6	0.56	0.54	0.48	0.56	0.28	97.58	97.459	2.52	0.021
DD&N	500	6	0.74	0.44	0.46	0.42	0.34	97.6	97.388	2.591	0.021

Table 7: Relative frequencies of strategies and PD Game outcomes averaged over the last 10 generations. AC denotes *Always Cooperate*, DD denotes *Defect Defectors*, and DD&N denotes *Defect Defectors and Neutrals*. Independent parameters are in the columns with gray heads.

As can be seen in Table 7, when the population size was large ($n = 500$) and when the *AlwaysCooperate* rule was used, the only possible outcome was that *thugs* invaded the population. When the population size was smaller, the results appeared to be random, so we ran the simulation for the second time, varying the following parameters (Table 8):

Parameter	Used values
population	{50, 100}
deterrence	{1, 3, 6}
rule	{ <i>DefectD</i> , <i>DefectD&N</i> }

Table 8: Parameters varied in the second run of Model 2. Each combination was run 10 times. That gave $10 * 2 * 3 * 2 = 120$ runs in total. All simulations were run for 500 generations.

We then counted the number of simulations for each combination of parameters invaded by *ranger* and *thug* agents. These results are presented in Table 9.

Population	Deterrence	Rule	Rangers	Thugs
50	1	DD	6	4
50	1	DD&N	6	4
50	3	DD	3	7
50	3	DD&N	6	4
50	6	DD	6	4
50	6	DD&N	8	2
100	1	DD	9	1
100	1	DD&N	9	1
100	3	DD	5	5
100	3	DD&N	5	5
100	6	DD	7	3
100	6	DD&N	8	2

Table 9: Parameters varied in the second run of Model 2. Each combination was run 10 times. That gave $10 * 2 * 3 * 2 = 120$ runs in total. All simulations were run for 500 generations.

We also decided to change the assumption that thugs are born with the estimate equal to 0.9 (see Appendix C) and observe if giving them even more advantage at the beginning will lead to more invasions. We ran our simulation once again varying the parameters as indicated in Table 8, but this time setting initial estimate of thugs to 0.1. It turned out that this does not significantly affect results and leads to similar pattern as depicted in Table 9 (63.33% invaded by rangers if initial estimate set to 0.1 vs. 65% with initial estimate set to 0.9).

Notes

¹The strength of agents is distributed normally within the population (mean = 10, SD = 1).

²At the beginning of every generation this list is empty – agents only start to learn through their experience and observation.

³Bone et al. (2015) criticize previous models for exactly the same reason and explicitly introduce power asymmetries. Nevertheless, they still assume that weak players are able to punish strong ones, albeit inflicting smaller relative damage.

⁴For example, *punishment* is roughly the same as benefit from the defection, *submit_cost* is smaller, because this is the case of accepting defeat to avoid fighting, etc.

⁵Note that when deterrence parameter is set to 0, values of other parameters should become irrelevant – that was indeed the case – in these simulations, we did not observe any variability in CC rates across combinations of other parameters.

⁶In the first generation, thugs are born with a belief that they are in majority; this gives them a slight advantage as they defect in the first round. This is motivated by our assumption that they should feel entitled to exploit others by default – they will only stop doing so if they learn they are in minority. In all subsequent generations they are born with the average belief present in the previous generation (see Appendix C for details).

⁷This is a generalization of the Grim Trigger strategy (Friedman 1971) that is even more grim, since justified defections will follow after merely observing the other player playing defect.

⁸In case of group conflicts, sensitivity used in decision to fight (Equation 1, Appendix A) needs to be lower as strength differences taken as input increase (see Appendix C for explanation).

⁹Halevy & Halali (2015) investigate third-party interventions, but in their framework benefits are introduced directly to the payoff structure, which allows strategic rather than morally-motivated interventions.

References

- Abbink, K. & de Haan, T. (2014). Trust on the brink of Armageddon: The first-strike game. *European Economic Review*, 67, 190–196
- Abbink, K. & Herrmann, B. (2011). The moral costs of nastiness. *Economic Inquiry*, 49(2), 631–633
- Abbink, K. & Sadrieh, A. (2009). The pleasure of being nasty. *Economics Letters*, 105(3), 306–308
- Alexander, R. D. (1987). *The Biology of Moral Systems*. Hawthorne, NY: Aldine de Gruyter
- Alexander, R. D. (1993). Biological considerations in the analysis of morality. In M. H. Nitecki & D. V. Nitecki (Eds.), *Evolutionary Ethics*, (pp. 163–196). Albany, NY: State University of New York Press
- Andrighetto, G., Giardini, F. & Conte, R. (2013). An evolutionary account of reactions to a wrong. In M. Knauff, M. Pauen, N. Sebanz & I. Wachsmuth (Eds.), *Proceedings of the 35th Annual Conference of the Cognitive Science Society*, (pp. 1756–1761). Austin, TX: Cognitive Science Society
- Arnott, G. & Elwood, R. W. (2009). Assessment of fighting ability in animal contests. *Animal Behaviour*, 77(5), 991–1004
- Asao, K. & Buss, D. M. (2016). The tripartite theory of Machiavellian morality: Judgment, influence, and conscience as distinct moral adaptations. In T. K. Shackelford & R. D. Hansen (Eds.), *The Evolution of Morality*, (pp. 3–25). Cham: Springer International
- Axelrod, R. & Hamilton, W. D. (1981). The evolution of cooperation. *Science*, 211(4489), 1390–1396
- Bandini, S., Manzoni, S. & Vizzari, G. (2009). Agent based modeling and simulation: An informatics perspective. *Journal of Artificial Societies and Social Simulation*, 12(4), 4
- Barker, J. L. & Barclay, P. (2016). Local competition increases people's willingness to harm others. *Evolution and Human Behavior*, 37(4), 315–322
- Batson, C. D. (2016). *What's Wrong with Morality? A Social-Psychological Perspective*. New York, NY: Oxford University Press
- Baumard, N., André, J.-B. & Sperber, D. (2013). A mutualistic approach to morality: The evolution of fairness by partner choice. *Behavioral and Brain Sciences*, 36(1), 59–78
- Benard, S. (2013). Reputation systems, aggression, and deterrence in social interaction. *Social Science Research*, 42(1), 230–245
- Benard, S. (2015). The value of vengeance: Reputational incentives for initiating versus reciprocating aggression. *Rationality and Society*, 27(2), 129–160
- Bone, J. E., Wallace, B., Bshary, R. & Raihani, N. J. (2015). The effect of power asymmetries on cooperation and punishment in a prisoner's dilemma game. *PloS ONE*, 10(1), e0117183
- Boyd, R. & Mathew, S. (2015). Third-party monitoring and sanctions aid the evolution of language. *Evolution and Human Behavior*, 36(6), 475–479
- Burnham, T. C. & Johnson, D. D. P. (2005). The biological and evolutionary logic of human cooperation. *Analyse & Kritik*, 27(1), 113–135
- Carlsmith, K. M. & Darley, J. M. (2008). Psychological aspects of retributive justice. *Advances in Experimental Social Psychology*, 40, 193–236
- Carlsmith, K. M., Darley, J. M. & Robinson, P. H. (2002). Why do we punish? deterrence and just deserts as motives for punishment. *Journal of Personality and Social Psychology*, 83(2), 284–299
- Charness, G., Masclet, D. & Villeval, M. C. (2013). The dark side of competition for status. *Management Science*, 60(1), 38–55
- Chase, I. D., Tovey, C., Spangler-Martin, D. & Manfredonia, M. (2002). Individual differences versus social dynamics in the formation of animal dominance hierarchies. *Proceedings of the National Academy of Sciences of the United States of America*, 99(8), 5744–5749

- Clutton-Brock, T. (2009). Cooperation between non-kin in animal societies. *Nature*, 462(7269), 51–57
- Clutton-Brock, T. H. & Parker, G. A. (1995). Punishment in animal societies. *Nature*, 373(6511), 209–216
- Crockett, M. J., Özdemir, Y. & Fehr, E. (2014). The value of vengeance and the demand for deterrence. *Journal of Experimental Psychology: General*, 143(6), 2279
- Cronk, L. (1994). Evolutionary theories of morality and the manipulative use of signals. *Zygon*, 29(1), 81–101
- DeScioli, P. (2016). The side-taking hypothesis for moral judgment. *Current Opinion in Psychology*, 7, 23–27
- DeScioli, P. & Kurzban, R. (2009). Mysteries of morality. *Cognition*, 112(2), 281–299
- DeScioli, P. & Kurzban, R. (2013). A solution to the mysteries of morality. *Psychological Bulletin*, 139(2), 477–496
- dos Santos, M. & Wedekind, C. (2015). Reputation based on punishment rather than generosity allows for evolution of cooperation in sizable groups. *Evolution and Human Behavior*, 36(1), 59–64
- Dreber, A. & Rand, D. G. (2012). Retaliation and antisocial punishment are overlooked in many theoretical models as well as behavioral experiments. *Behavioral and Brain Sciences*, 35(1), 24
- Fawcett, T. W. & Johnstone, R. A. (2010). Learning your own strength: Winner and loser effects should change with age and experience. *Proceedings of the Royal Society of London B: Biological Sciences*, (p. rspb20092088)
- Fawcett, T. W. & Mowles, S. L. (2013). Assessments of fighting ability need not be cognitively complex. *Animal Behaviour*, 86(5), e1–e7
- Fehr, E. & Fischbacher, U. (2004). Third-party punishment and social norms. *Evolution and Human Behavior*, 25(2), 63–87
- Fehr, E. & Gächter, S. (2002). Altruistic punishment in humans. *Nature*, 415(6868), 137–140
- Franz, M., McLean, E., Tung, J., Altmann, J. & Alberts, S. C. (2015). Self-organizing dominance hierarchies in a wild primate population. *Proceedings of the Royal Society of London B: Biological Sciences*, 282(1814), 20151512
- Friedman, J. W. (1971). A non-cooperative equilibrium for supergames. *The Review of Economic Studies*, 38(1), 1–12
- Garcia, M. J., Murphree, J., Wilson, J. & Earley, R. L. (2014). Mechanisms of decision making during contests in green anole lizards: Prior experience and assessment. *Animal Behaviour*, 92, 45–54
- Gardner, A. & West, S. A. (2004). Cooperation and punishment, especially in humans. *The American Naturalist*, 164(6), 753–764
- Giardini, F., Andrighetto, G. & Conte, R. (2010). A cognitive model of punishment. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*, (pp. 1282–1288). Austin, TX: Cognitive Science Society
- Giardini, F. & Conte, R. (2015). Revenge and conflict: Social and cognitive aspects. In F. D’Errico, I. Poggi, A. Vinciarelli & L. Vincze (Eds.), *Conflict and Multimodal Communication: Social Research and Machine Intelligence*, (pp. 71–89). Cham: Springer International
- Gilbert, N., den Besten, M., Bontovics, A., Craenen, B. G. W., Divina, F., Eiben, A. E., Griffioen, R., Hévízi, G., Lőrincz, A., Paechter, B., Schuster, S., Schut, M. C., Tzolov, C., Vogt, P. & Yang, L. (2006). Emerging artificial societies through learning. *Journal of Artificial Societies and Social Simulation*, 9(2), 9
- Gordon, D. S., Madden, J. R. & Lea, S. E. G. (2014). Both loved and feared: Third party punishers are viewed as formidable and likeable, but these reputational benefits may only be open to dominant individuals. *PLoS ONE*, 9(10), e110045
- Griskevicius, V., Tybur, J. M., Gangestad, S. W., Perea, E. F., Shapiro, J. R. & Kenrick, D. T. (2009). Aggress to impress: Hostility as an evolved context-dependent strategy. *Journal of Personality and Social Psychology*, 96(5), 980–994
- Hagen, E. H. & Hammerstein, P. (2006). Game theory and human evolution: A critique of some recent interpretations of experimental games. *Theoretical Population Biology*, 69(3), 339–348

- Halevy, N. & Halali, E. (2015). Selfish third parties act as peacemakers by transforming conflicts and promoting cooperation. *Proceedings of the National Academy of Sciences of the United States of America*, 112(22), 6937–6942
- Helbing, D., Szolnoki, A., Perc, M. & Szabó, G. (2010). Evolutionary establishment of moral and double moral standards through spatial interactions. *PLoS Computational Biology*, 6(4), e1000758
- Herrmann, B., Thöni, C. & Gächter, S. (2008). Antisocial punishment across societies. *Science*, 319(5868), 1362–1367
- Hick, K., Reddon, A. R., O'Connor, C. M. & Balshine, S. (2014). Strategic and tactical fighting decisions in cichlid fishes with divergent social systems. *Behaviour*, 151(1), 47–71
- Hobbes, T. (2011 [1651]). *Leviathan: Or the matter, forme and power of a commonwealth ecclesiastical and civil*. London: Penguin
- Houser, D. & Xiao, E. (2010). Inequality-seeking punishment. *Economics Letters*, 109(1), 20–23
- Hsu, Y., Earley, R. L. & Wolf, L. L. (2006). Modulation of aggressive behaviour by fighting experience: Mechanisms and contest outcomes. *Biological Reviews*, 81(1), 33–74
- Jaffe, K. & Zaballa, L. (2010). Co-operative punishment cements social cohesion. *Journal of Artificial Societies and Social Simulation*, 13(3), 4
- Jensen, K. (2010). Punishment and spite, the dark side of cooperation. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 365(1553), 2635–2650
- Johnson, D. D. P. & Fowler, J. H. (2011). The evolution of overconfidence. *Nature*, 477(7364), 317–320
- Johnstone, R. A. & Bshary, R. (2004). Evolution of spite through indirect reciprocity. *Proceedings of the Royal Society of London B: Biological Sciences*, 271(1551), 1917–1922
- Jordan, J. J., Hoffman, M., Bloom, P. & Rand, D. G. (2016). Third-party punishment as a costly signal of trustworthiness. *Nature*, 530(7591), 473–476
- Krasnow, M. M., Cosmides, L., Pedersen, E. J. & Tooby, J. (2012). What are punishment and reputation for? *PLoS ONE*, 7(9), e45662
- Krasnow, M. M., Delton, A. W., Cosmides, L. & Tooby, J. (2015). Group cooperation without group selection: Modest punishment can recruit much cooperation. *PLoS ONE*, 10(4), e0124561
- Krasnow, M. M., Delton, A. W., Cosmides, L. & Tooby, J. (2016). Looking under the hood of third-party punishment reveals design for personal benefit. *Psychological Science*, 27(3), 405–418
- Kurzban, R. & DeScioli, P. (2013). Adaptationist punishment in humans. *Journal of Bioeconomics*, 15(3), 269–279
- Lahti, D. C. & Weinstein, B. S. (2005). The better angels of our nature: Group stability and the evolution of moral tension. *Evolution and Human Behavior*, 26(1), 47–63
- Lammers, J., Stapel, D. A. & Galinsky, A. D. (2010). Power increases hypocrisy: Moralizing in reasoning, immorality in behavior. *Psychological Science*, 21(5), 737–744
- Marczyk, J. (2015). Moral alliance strategies theory. *Evolutionary Psychological Science*, 1(2), 77–90
- Marlowe, F. W., Berbesque, J. C., Barrett, C., Bolyanatz, A., Gurven, M. & Tracer, D. (2011). The 'spiteful' origins of human cooperation. *Proceedings of the Royal Society of London B: Biological Sciences*, (p. rspb20102342)
- Maze, J. R. (1973). The concept of attitude. *Inquiry*, 16(1–4), 168–205
- McCullough, M. E., Kurzban, R. & Tabak, B. A. (2013). Cognitive systems for revenge and forgiveness. *Behavioral and Brain Sciences*, 36(1), 1–15
- Miller, D. T. (2001). Disrespect and the experience of injustice. *Annual Review of Psychology*, 52(1), 527–553
- Nakao, H. & Machery, E. (2012). The evolution of punishment. *Biology & Philosophy*, 27(6), 833–850
- Nietzsche, F. W. (2006 [1887]). *On the Genealogy of Morality*. Cambridge: Cambridge University Press

- Nowak, A., Gelfand, M. J., Borkowski, W., Cohen, D. & Hernandez, I. (2016). The evolutionary basis of honor cultures. *Psychological Science*, 27(1), 12–24
- Okada, D. & Bingham, P. M. (2008). Human uniqueness: Self-interest and social cooperation. *Journal of Theoretical Biology*, 253(2), 261–270
- Olson, M. (1971 [1965]). *The Logic of Collective Action: Public Goods and the Theory of Groups*. Cambridge, MA: Harvard University Press
- Petersen, M. B., Sznycer, D., Sell, A., Cosmides, L. & Tooby, J. (2013). The ancestral logic of politics: Upper-body strength regulates men's assertion of self-interest over economic redistribution. *Psychological Science*, 24(7), 1098–1103
- Pietraszewski, D. & Shaw, A. (2015). Not by strength alone. *Human Nature*, 26(1), 44–72
- Prenter, J., Elwood, R. W. & Taylor, P. W. (2006). Self-assessment by males during energetically costly contests over precopula females in amphipods. *Animal Behaviour*, 72(4), 861–868
- Price, M. E., Cosmides, L. & Tooby, J. (2002). Punitive sentiment as an anti-free rider psychological device. *Evolution and Human Behavior*, 23(3), 203–231
- Przepiorka, W. & Liebe, U. (2016). Generosity is a sign of trustworthiness – the punishment of selfishness is not. *Evolution and Human Behavior*, 37(4), 255–262
- Rabellino, D., Morese, R., Ciaramidaro, A., Bara, B. G. & Bosco, F. M. (2016). Third-party punishment: Altruistic and anti-social behaviours in in-group and out-group settings. *Journal of Cognitive Psychology*, 28(4), 486–495
- Raihani, N. J. & Bshary, R. (2015). Why humans might help strangers. *Frontiers in Behavioral Neuroscience*, 9(39), 1–11
- Schelling, T. C. (1960). *The Strategy of Conflict*. Cambridge, MA: Harvard University Press
- Sell, A., Eisner, M. & Ribeaud, D. (2016). Bargaining power and adolescent aggression: The role of fighting ability, coalitional strength, and mate value. *Evolution and Human Behavior*, 37(2), 105–116
- Sell, A., Hone, L. S. & Pound, N. (2012). The importance of physical strength to human males. *Human Nature*, 23(1), 30–44
- Sell, A., Tooby, J. & Cosmides, L. (2009). Formidability and the logic of human anger. *Proceedings of the National Academy of Sciences of the United States of America*, 106(35), 15073–15078
- Sheskin, M., Bloom, P. & Wynn, K. (2014). Anti-equality: Social comparison in young children. *Cognition*, 130(2), 152–156
- Shoham, Y. & Tennenholtz, M. (1997). On the emergence of social conventions: Modeling, analysis, and simulations. *Artificial Intelligence*, 94(1), 139–166
- Sylwester, K., Herrmann, B. & Bryson, J. J. (2013). Homo homini lupus? Explaining antisocial punishment. *Journal of Neuroscience, Psychology, and Economics*, 6(3), 167
- Tooby, J. & Cosmides, L. (2010). Groups in mind: The coalitional roots of war and morality. In H. Høgh-Olesen (Ed.), *Human Morality and Sociality: Evolutionary and Comparative Perspectives*, (pp. 191–234). New York, NY: Palgrave MacMillan
- Trivers, R. L. (1971). The evolution of reciprocal altruism. *The Quarterly Review of Biology*, 46(1), 35–57
- Villatoro, D., Andrighetto, G., Brandts, J., Nardin, L. G., Sabater-Mir, J. & Conte, R. (2014). The norm-signaling effects of group punishment: Combining agent-based simulation and laboratory experiments. *Social Science Computer Review*, 32(3), 334–353
- West, S. A., El Mouden, C. & Gardner, A. (2011). Sixteen common misconceptions about the evolution of cooperation in humans. *Evolution and Human Behavior*, 32(4), 231–262
- West, S. A., Griffin, A. S. & Gardner, A. (2007). Social semantics: Altruism, cooperation, mutualism, strong reciprocity and group selection. *Journal of Evolutionary Biology*, 20(2), 415–432