Understanding and Addressing Cultural Variation in Costly Antisocial Punishment

Joanna J. Bryson[1], James Mitchell[1], Simon T. Powers[2], and Karolina Sylwester[1]

[1] Artificial Models of Natural Intelligence
Department of Computer Science
University of Bath
Bath, BA2 7AY
England, United Kingdom
J.J.Bryson@bath.ac.uk
[2] Department of Ecology & Evolution
University of Lausanne
CH-1015 Lausanne
Switzerland
Simon.Powers@Unil.ch

Abstract. Altruistic punishment — punishment of those contributing little to the pubic good — has been proposed as an explanation for the extraordinary extent of human culture relative to other species. culture. Altruistic punishment is seen as supporting the high levels of altruism necessary for the cooperation underlying this culture, including information exchange. However, humans will also sometimes punish those who contribute to the public good, even when those contributions directly benefit the punisher. This behaviour — antisocial punishment — is negatively correlated with GDP, and as such may be seen as a hinderance to overall wellbeing. In this chapter, we pursue a better understanding of antisocial punishment in particular and costly punishment in general. We explore existing data showing cultural variation in the propensity to punish, and ask how such sanctioning, whether of those who give much or little, affects the individuals who conduct it. We hypothesise that costly punishment is a mechanism for regulating investment between different levels of society, for example whether an individual's current focus should be on their nation, village, family or self. We suggest that people are less likely to antisocially punish those they consider to be "in group", and that the propensity to apply this identity to strangers may vary with socio-economic-political context and resulting individual experience. In particular, an increased propensity to express antisocial punishment should correlate with a lower probability of benefiting from public goods, as may be the case where there is low rule of law. We show both analysis of behavioural economics experiments and evolutionary social simulations to support our hypotheses, and suggest implications for policy makers and other organisations that may wish to intervene to improve general economic wellbeing.

Keywords: antisocial punishment (ASP); altruistic punishment (AP); costly punishment; public goods; public goods games (PGG); behavioural economics; altruism; cooperation; in-group / out-group assessment.

1 Introduction

> That friendship lasts longest—if there is a chance of its being a success—
> in which friends both give and receive gifts. — *The Hávamál*[3].

The variety of human cultures is one of the joys of contemporary human life. However, a respect and appreciation for diversity cannot be allowed to mask the observation that cultural variation can include measurable differences in metrics that have nearly-universal cross-cultural appeal, for example reducing infant mortality or increasing literacy. For the last several years we have been striving to understand cultural variation in one such trait: the propensity of individuals to optimise economic collaboration when thrown into a group together. In this chapter we review our progress to date. We also examine the policy implications of our findings on cooperation and punishment, particularly for organisations wishing to aid development or rebuild communities in areas experiencing conflict.

The behaviour we are studying is called *anti-social punishment* (ASP). Technically, ASP occurs when an individual is willing to pay a penalty to punish a member of their own group, where the victim of the punishment has been generous, providing *more or equal* contributions to their mutual group than the punisher. The term *punishment* here is being used to describe a punisher deliberately paying a cost to have money taken away from the the victim of their punishment. For the data analysed here, this is all done anonymously in an experimental context, with the experimenter acting as the go-between, executing the instructions of the participants. In ordinary experience, we believe this behaviour to map to the situation where generosity or philanthropy is rejected, resented or punished. In sociology, this relates to the theory of the gift, where healthy equal exchange increases bonds, but gifts that cannot or will not be reciprocated are seen as a power move, an extreme version of which is potlatch (Mauss, 1967).

Systematic cultural variation in ASP behaviour was first documented in the economics literature by Herrmann et al. (2008a), and Benedikt Herrmann has been one of our collaborators throughout this project. Although the data Herrmann et al. provide is based on formal laboratory experiments where participants play a 'game' for money, the results correlate highly with national Gross Domestic Product (GDP), suggesting the possibility that the behaviour measured in the laboratory may have fundamental impact on the economic wellbeing of a nation, though of course the reverse could also be true. Further, the variation between cultures is not arbitrary, but rather appears clustered by global region. Thus Boston, several cities in Northern Europe, the Far East, and Melbourne show high levels of profitable economic collaboration, while Athens, Istanbul, regions of the Middle East and of the former Soviet Union show relatively low ability to collaborate for profit, and higher levels of ASP.

If we can find an explanation for such variation, we might not only be rewarded with a better understanding of culture more generally. We might also

[3] Translated by Martin Clarke (1923), also quoted by Mauss (1967).

be in a better situation to administer economic aid, or to otherwise shape intervention policies. If generosity is perceived as a power move to be resented and, if possible, rejected, then clearly it is less likely to be effective. How can we bring interventions to be perceived as a collaborative effort to mutually improve economies and / or security? In the present chapter our focus is at the level of the city and state, but there is also relevance to managing interpersonal relationships and individual socialisation and well being.

This chapter begins with a review of the scientific context of our research. We then review our findings, some of which have been published previously, others of which are presented here for the first time. Overall, we have failed to find any evolutionary context in which ASP can evolve unless we assume that it carries some extra benefit beyond its economic costs. We hypothesise that this benefit is social status awarded to those who punish. In the results given here, we model the simplest case, which is awarding status regardless of whether the punishment is altruistic (punishing those donating less than the punisher to the group) or antisocial[4]. If we include this assumption, then we *are* able to account for variation in ASP. We suggest that regional variation in ASP reflects the extent to which in various societies one's wellbeing depends on one's relative status within one's own group rather than the group's status in relation to other groups. After reviewing these findings, we discuss policy implications for our work. We make a number of suggestions, then close with our conclusions.

2 Scientific Background: Costly Punishment

Herrmann et al. (2008a) show that in some human subject pools (e.g. university undergraduates in Boston, Melbourne, Chengdu and Zurich) group members tend to quickly exploit an experimental context in which mutual investment leads to mutual benefits. However, in other societies (e.g. university undergraduates in Muscat, Istanbul, Minsk and Athens) substantial proportions of participants will pay a fee in order to penalize group members more generous to the group's public good than themselves. This is despite the fact that this generosity is benefiting all group members *other than* the benefactor, including the punisher's. Such punishment of cooperation is called *antisocial punishment* (ASP).

Herrmann et al. sought correlates for the prevalence of ASP in a culture, finding several. ASP inversely correlates with both Gross Domestic Product (GDP) and the Rule of Law (Kaufmann et al., 2004). They suggest that "weak norms of civic cooperation and the weakness of the rule of law in a country are significant predictors of antisocial punishment. Our results show that [...] punishment opportunities are socially beneficial only if complemented by strong social norms of cooperation." But correlation does not demonstrate causation. Can we be sure that the propensity for ASP does not itself lead to a weak rule of law? Or that both could be caused by some other factor? In the next section, we describe the data in greater detail, and try to answer these questions.

[4] Of course, reality could be more subtle (Barclay, 2006; Sylwester et al., 2013a).

2.1 The Data: How Cooperation and Punishment Are Measured

All human subject data for this research was collected using a paradigm from a relatively new branch of economics, called either *experimental economics* or *behavioural economics*. Behavioural economics is similar to experimental psychology, except that economists mandate certain conditions, for example all individuals must receive sufficient financial reward to be considered motivated. Significantly, subjects can in no way be deceived, and must in fact demonstrate understanding of the complete consequences of their own and the other team members' possible actions by passing a test before participating in the games. Similarly, in cross-cultural experimental economics research, the players play for tokens, to keep reasoning about proportions equally easy for subjects regardless of local currency denominations (Herrmann et al., 2008b).

The standard behavioural economics experiment for assessing costly punishment is called the Public Goods Game (PGG, Ledyard, 1995). In the basic form of this game there is no punishment. In the standard form, a group is determined by an experimenter, but members are not identified to each other and only interact by computer screens[5]. This anonymity simplifies theoretical analysis, by ensuring that group members do not act out of fear or expectation of retribution or reward after the game. In a single round of PGG, each member is given 20 tokens by the experimenter. Subjects are then allowed to contribute any portion they choose of their allocation to the public pool. Allocations to the public pool are multiplied by the experimenter then divided equally between all group members. The multiplication factor is always greater than one but smaller than the number of group members. As a result, the optimal outcome for the group as a whole is for all to contribute everything, but *individual* investments are never fully returned. For example, if the multiplier is 3 and the number of individuals is 4, then for every token an individual donates, they (and every other member of the group) receive 3/4 of a token back. Thus individuals who do not contribute anything or contribute less than others gain a financial advantage relative to those others, at least for that round. A PGG therefore represents a social dilemma because an individual's interests are in conflict with their group's. In the experiments described below, PGGs are played repeatedly, for ten consecutive rounds with the same groups.

In the punishment condition, after a round of PGG individuals can anonymously punish others. The target to be punished can only be identified only by their previous round's contribution to the public pool. Importantly, subjects never learn any information about who punishes them, only the size of their most recent contribution. Punishment is costly; in the studies described here, for every token a punisher pays, the punishee loses three tokens[6]. When an individual punishes someone who has contributed less than they have, this punishment is

[5] In rural conditions the computers may be replaced with pen and paper for recording decisions, then the results are communicated to group members by the experimenter.

[6] Many other cost/effect ratios have been tried by other experimenters, these result in quantitative but not qualitative shifts in behaviour. See (Sylwester et al., 2013a) for a more complete review.

termed *altruistic* (AP) because the punisher pays a cost, yet the whole group benefits if (as seems often to be the case) this action leads to higher contributions. On the other hand, if punishers punish those who contribute more or equally to themselves, the punishment is called *anti-social*. Herrmann et al. (2008a) were the first to document societies with large amounts of ASP, and showed that this could in some cases completely counter the expected benefits, in fact reducing overall cooperation and payoffs to the subjects. In the Swiss contexts where these experiments were first run, the opportunity to punish reliably resulted in a better economic outcome for subjects playing the PGG. However, this was not true in some societies with high levels of ASP. In most of the data reported here (all of which is due to Herrmann et al.) subjects played two rounds of 10 PGG, one with punishment and one without. For most subject pools the order of the games (punishment or not) was randomised.

2.2 Previous Interpretations of Punishment Results

To fully understand the literature and history of work in costly punishment, we must recognise that one goal of anthropology is explaining human uniqueness. Why are humans the only species with advanced technology? Why are we dominating the biomass of the planet with our ever-expanding population? The explanation is not simple biology — it is not just our intelligence or our capacity for tool use. The vast majority of population growth and technological complexity is of very recent origin. Very human-like species existed and used primitive tools for millions of years (Walker and Stringer, 2010). Urbanisation, agriculture, writing and doctrinal religions (those that share their practices outside of small close-knit tribal structures) all seem to date to no more than 8,000-12,000 years ago, well after the first appearance of *Homo sapiens*.

Numerous empirical and theoretical studies have suggested or proposed an extraordinary human propensity for cooperation as an explanation for the extent of human culture (e.g. Gintis et al., 2003; Henrich et al., 2001). However, the reasoning could just as easily be applied in reverse; it could be that a an extraordinary human propensity for accumulating culture accounts for the extent of human cooperation (Bryson, 2009). Thus the extent of human cooperation is not yet considered fully accounted for. After the early PGG punishment results (e.g. Fehr and Gächter, 2000; Fehr and Gächter, 2002), altruistic punishment was regarded as a possible explanation for large-scale cooperation. Here too though the reasoning seemed cyclic, as punishment can be a form of cooperation itself, and contrary to its reputation, altruistic behaviour is neither difficult to evolve nor uniquely human (Čače and Bryson, 2007; West et al., 2007; MacLean et al., 2010). Swinging to the other extreme, the phenomenon of ASP has more recently lead some scientists to emphasise the 'dark side' of human behaviour, including a tendency for spite and hyper-competitiveness (Abbink and Sadrieh, 2009; Jensen, 2010; Sylwester et al., 2013a). Extremes of moral assessment and defensiveness need to be guarded against if we are to understand what underlies these phenomenon. We believe the Herrmann et al. (2008a) results indicate that punishment is part of a much more complex system of social regulation,

not a simple explanation for human cooperation and therefore culture. Here we attempt to approach the explanation of costly punishment objectively, by viewing both cooperation and punishment as biological phenomena and looking for ultimate causes that might make such behaviour adaptive. In the next section, we briefly review the sorts of explanations natural selection can provide for behaviour.

2.3 Proximate and Ultimate Explanations

Fields like evolutionary anthropology and behavioural ecology work from the assumption that behaviour is inseparable from the rest of the organism, including in terms of its causal explanation. From the perspective of evolution there are at least two types of causes for any trait (Mayr, 1961). *Ultimate causes* concern why the behaviour is present in a population as whole — what role does it serve in the evolutionary struggle? Contemporary evolutionary theory does not expect all observed traits to be adaptations — some are incidental side-effects of historical associations, since the selection process takes time and can only operate on the material at hand. Nevertheless, it is at least a common first guess in evolutionary approaches that an observed trait exists because it has historically provided more advantage than disadvantage to those who hold it relative to those that do not. *Proximate causes* in contrast describe mechanism — what triggers and/or enables a particular organism to perform the behaviour in question. For example, running may be ultimately a good way to escape, and proximately a response to a loud noise. Note that for some species, flying or swimming is a better mode of escape than running. Identifying the ultimate cause of a behaviour does not mean that behaviour is necessarily the optimal mechanism for meeting that need. Which behaviour will be expressed also depends on evolutionary (phylogenetic) history.

A useful proximate mechanism may itself become an ultimate explanation for some other trait. For example, hearing the sound of a predator may be the proximate cause of fleeing, *and* the ultimate cause for large ears. This sort of complexity has lead some to suggest that the distinction between these causes is not real and creates an impediment to understanding. (e.g. Thierry, 2005; Laland et al., 2011). However, the distinction was developed to address a still-common error — thinking that a simple proximate explanation for a behaviour can displace a complex ultimate one. While science generally favours simpler explanations, because proximate and ultimate explanations address different questions one cannot substitute for the other.

3 Building an Understanding of Anti-Social Investment

In this section we introduce our findings concerning an explanation for the behaviour termed *anti-social punishment*. We begin by explaining our current hypothesis, then review the evidence we have discovered leading us to this hypothesis.

3.1 Hypothesis: Punishment as Regulation

Our current hypothesis is that all punishment is an aggressive act, which some proportion of any population is motivated to perform. A proximate reward for aggression is increased social status, not only relative to the target of the aggressive act but also to bystanders who witness the aggression. When an individual becomes known to be aggressive, confronting that individual becomes associated with an increased cost, thus making avoiding confrontation via submissive gestures a more attractive strategy (Preuschoft and van Schaik, 2000). We also hypothesise, however, that in contexts where cooperation is more likely to produce stable public goods, members of the population are also more likely inhibit any tendency they might feel to be aggressive towards cooperators. We think that for at least some proportion of the population, whether cooperative gestures are accepted as useful or seen as another form of dominance / aggression depends on whether the generous individual is seen as a member of a trusted "in group", or is seen as"out-group" — a potential invader. Thus the proximate explanation for a population with relatively high ASP is a larger number of individuals assessing anonymous strangers playing a PGG as out-group, and the ultimate explanation is a dependency on the expected utility of public goods in that population's socio-political-economic context. This expected utility is estimated by the individuals composing the population based on data from the experience of their lives up until the experiment, as interpreted through prior expectations communicated to those individuals by their culture, which reflects the experience of many more individuals.

Our hypothesised ultimate benefit derives from the observation that an investment in a global public good comes at a cost of reduced investment not only directly to the individual, but to other more-local goods such as the individual's family. The ultimate need to support many levels of investment may explain the otherwise odd tendency of nearly all subjects to split their investment strategy, keeping some proportion of the resources originally allocated to them by the experimenter, and investing the rest. When Northern Europeans (including Boston in the Herrmann et al. data) read the instructions concerning the punishment condition of the PGG, their expected utility allocated to the public good immediately increases. Interestingly, in the three cities tested in Australia and Asia the initial expectation seems to be the same in both conditions, but that expectation rises over the course of the multi-round game in response to increased revenue as rounds are played. Whereas for Athens, Istanbul, the Middle East and the former Soviet Union neither expectations nor rewards increase, and public goods investment stays approximately constant at a relatively low level throughout.

Thus at an ultimate level, we hypothesise that variation in punishment strategies may be an evolved mechanism for regulating global public goods investment (versus more local or individual investment) to a level appropriate to that population's economic context. What is appropriate is estimated in a distributed fashion by the population's individual experiences, and aggregated into a set of

collective norms and expectations that influence proximate responses to social dilemmas.

In the remainder of this section we review our evidence for this hypothesis. Where we describe human data results, these are derived from further analysis of the original Herrmann et al. (2008a). The additional analysis was performed mostly by Sylwester and Mitchell. However, we first turn to theoretical results derived from simulation. Simulation is a process of analysing the full consequences of a theory by describing it so thoroughly that it can be executed on a computer. All systems of modelling theory are analytic processes performed because, as Kokko and López-Sepulcre (2007) phrases it, "our brains aren't big enough" to see all of our theories' consequences. Thus for example, different researchers might dispute whether a minimalist theory is really sufficient to explain the complex behaviour observed in the real world. Formal modelling can demonstrate with certainty whether the results of a hypothesised system are as predicted, though it *cannot* determine with certainty whether this reflects what happens in the real world. Models are ultimately only theories, and their validity is assessed by standard scientific processes of assessing fit to data and evaluating unexpected predictions (see further Bryson et al., 2007; Whitehouse et al., 2012). Computer modelling (simulations) also allows us to check for internal coherence of our theories, since inconsistant theories are impossible to build and run as programs.

Most of the modelling performed here is a form of simulation known as agent-based modelling (ABM). That is, abstracted versions of both the individual actors (agents) and the environment in which they act are programmed into computers. The abstractions include the hypothesised minimal set of actions the agents are capable of, knowledge the agents must retain to inform this action, elements of the environment (including other agents) the agents may act on, and environmental results of these actions. A computer then executes the operation of the agents over time and reports the consequences. Where these are not completely deterministic, many experimental runs may be performed to discover the distribution of results. Where characteristic of the agents or environment are not known or believed to vary, again many runs can be performed with different values of these, to measure the consequences. Here, most of the described modelling has been performed by Powers and Taylor.

3.2 Ultimately, ASP Is Not Viable Unless It Correlates with Some Other Benefit

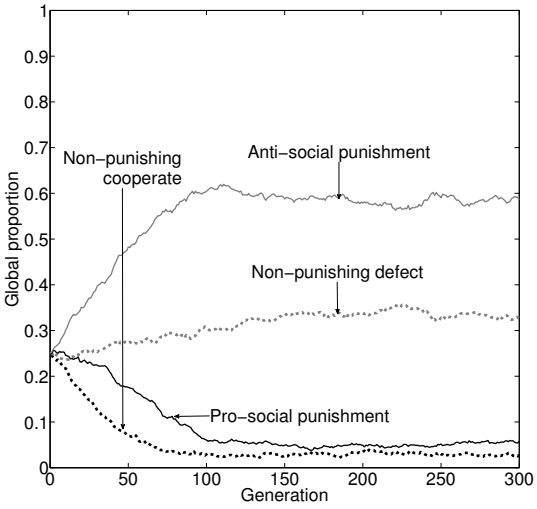
We begin our presentation of evidence for our hypotheses by examining the basic question of the circumstances in which punishment could have evolved as a strategy. The results presented here are based on multi-level evolutionary ABM. A multi-level model allows us to manipulate the relative costs and benefits of within-group and between-group competition. This is one way to think about local versus global (in the biological sense of less-local) investment and competition. Local competition occurs *within* the in group — for example, who in my family gets the biggest piece of pie? In contrast, global competition occurs

between groups — for example, which family gets the most pies? Note that there can be many levels of competition, and therefore selection. Families can join together to compete as one village against another; villages may join to compete as one state against another.

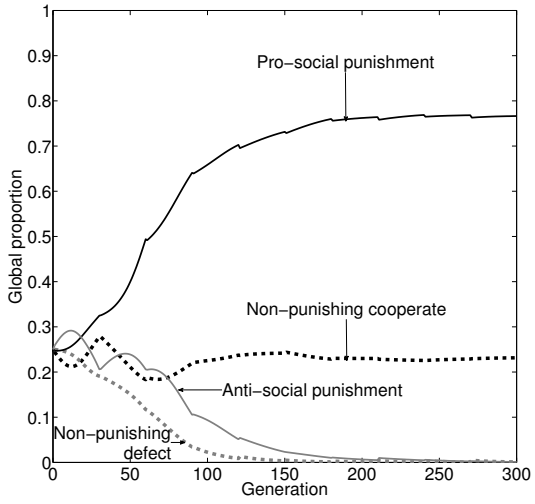
The multi-level ABM here extend from Powers et al. (2011). Within-group competition is increased by *increasing* the group size, since this increases the variance in social behaviour *within* groups, and so increases the strength of within-group selection. Between-group competition is likewise increased by *decreasing* group size, since this increases the variance in social behaviour *between* groups, and so increases the strength of between-group selection. The importance of between-group competition is also increased by *decreasing* the probability that individual agents find themselves in new groups, that is, by reducing the frequency with which groups are reformed. This may be thought of as modelling a decreased amount of communication and interdependency between groups in the real world, e.g. little intermarriage. Note though that both mechanisms serve as computationally-clear abstractions, and may represent more complex real-world variations in rewards at the different levels. For example, climate change could increase population pressure by reducing the amount of habitable territory. This could result in increased migration, changing the scale of competition from more local to more global.

Here we examine the viability of ASP in particular as a strategy, and also how its availability as a strategy affects the utility of costly punishment as a strategy over all. As in Powers et al. (2012), a linear public goods game with punishment is played within groups once per generation. The payoffs from this game determine the fitness of individuals, such that individuals with a high absolute payoff produce more offspring. Groups remain together for a particular number of generations. Then all individuals are considered a part of one global migrant pool, from which the next generation of groups is formed. This so-called *dispersal stage* creates between-group competition, since groups containing a larger number of individuals at the time of dispersal produce a larger fraction of the migrant pool, and hence have more impact in the next generation of groups. The size of a group at the time of dispersal is in turn affected by the mean payoff that its members receive from the public goods game.

In a thorough examination, Powers et al. could find no evolutionary context in which ASP was adaptive against other social strategies, unless we assume that punishment actually has a *negative* cost. That is, punishment must generate some benefit to the punisher in order for ASP to ever be adaptive. However, as explained earlier, punishment is definitionally costly and also, relative to other group members, altruistic, since any economic benefit accrued to the punisher due to e.g. an increase in PG investment is shared by the others even though they do not pay the cost of punishment. One example of how punishment might benefit the punisher despite costing risk of injury, effort and time, is if punishment takes the form of taking resources away from the target. If the punisher keeps these for themself rather than sharing with the rest of the group, this would compensate immediately for the risk of aggression (Taylor et al., 2013).



(a) Groups reform every generation.



(b) Groups reform after 30 generations.

Fig. 1: Evolution of strategy frequencies given that punishment provides direct benefit. (a) When groups reform regularly, within-group competition is the main driver of the evolutionary dynamics. Parameters: founding group size= 15, benefit from cooperation= 0.9, cost to cooperating= 0.1, cost of being punished= 0.3, cost of punishing= -0.1, groups randomly reformed every generation. (b) When groups stay together for multiple generations, between-group competition supports cooperative strategies. Parameters: As for a but with groups reforming every 30 generations.

However, even where the proximate outcome of punishment is fully public, there may be other longer-term benefits to the punisher, such as increased social status and its associated benefits (Preuschoft and van Schaik, 2000). We know that altruistic punishment in in-group contexts does lead to increased status (Barclay, 2006; Sylwester and Roberts, 2010).

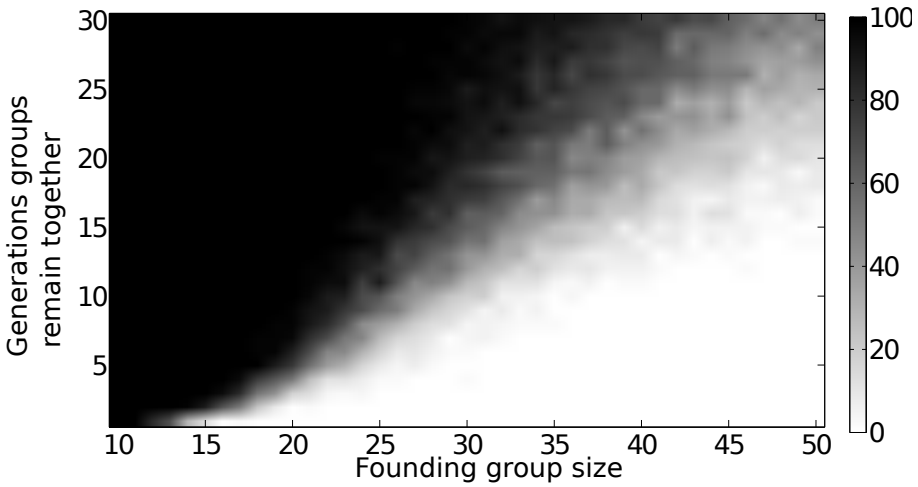
Our current guess is this latter option — that punishment is used to signal or even generate dominance within a group. The benefits of social dominance over the lifetime of the punisher may more than compensate for the immediate cost of the punishment act (West et al., 2011). Indeed, dominance is often seen as a form of long-term conflict resolution, because it reifies a particular set of expectations of conflict outcome, thus reducing the amount of actual physical conflict required (Preuschoft and van Schaik, 2000; Bryson et al., 2012). Thus, both AP and ASP may maintain or increase an individual's rank in a dominance hierarchy, which may in turn increase longe-term benefit and thus fitness relative to those who do not (Clutton-Brock and Parker, 1995; Boehm, 1999; Rohwer, 2007). But this guess has yet to be turned into a formal hypothesis, let alone tested. What we know from our simulations is only that *some* additional factor must account for the existence of ASP.

Even in the case where punishment *does* result in intrinsic benefit, then there is still an impact of local versus global competition. Where groups compete with each other — in the present ABM, where they persist long enough to receive fitness payoff for their public goods — prosocial (altruistic) punishment is still selected for over ASP (see Figure 1a). Only when within-group competition is the stronger selective force can even individually-advantageous ASP out compete the other form of punishment (Figure 1b).

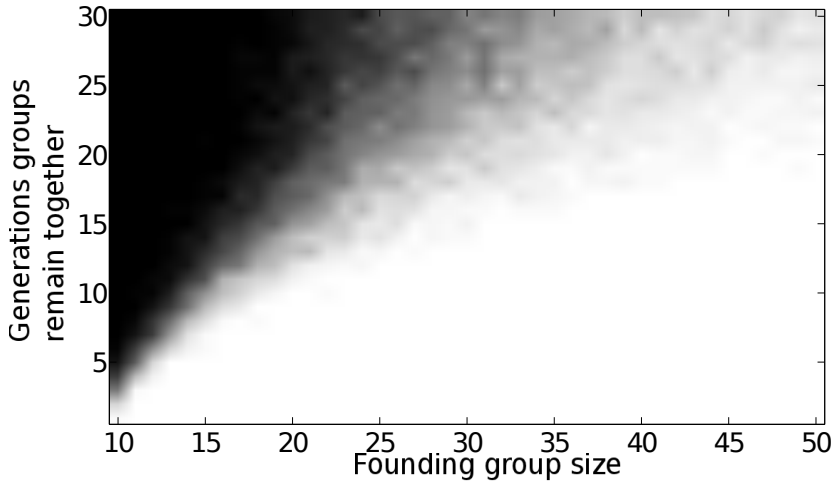
3.3 Punishment Alone Cannot Account for Human Sociality

As described in Section 2.2, the finding by Herrmann et al. (2008a), that populations exist in which the introduction of punishment reduces performance in PGG, was disruptive to those who believed that punishment explained exceptional levels of cooperation in humans. This result is sufficiently disruptive that it has been attacked on methodological grounds, either against behavioural economics in general or as practiced in the specific cases. However, modelling results show that even in pure theory, once ASP is taken into account, punishment cannot be considered solely a mechanism for increasing cooperation (Rand et al., 2010; Rand and Nowak, 2011; Powers et al., 2012).

As Figure 2a demonstrates, even where punishment is exclusively altruistic, cooperation will not necessarily be selected for. Only where group size is relatively small and relatively stable (there are many generations between dispersal stages) do cooperative strategies reliably evolve. This is because such conditions effectively increase the variance in social behaviour between groups (see Powers et al. 2012), and so create conditions for effective group selection. As a result, cooperation (including AP) is more likely to be a beneficial strategy and increase in prevalence. Conversely, a large founding group size and / or frequent group-mixing increases within-group variation in social behaviour, and hence makes



(a) Only altruistic punishment possible.



(b) Both altruistic and antisocial punishment possible.

Fig. 2: Percentage of Monte Carlo simulation runs in which pro-social punishment and cooperation together constituted more than 90% of the global population at equilibrium: (*a*) without the presence of anti-social punishment; (*b*) with anti-social punishment included. Note here (unlike Figure 1) punishment is assumed to be costly, thus ASP never dominates as a strategy, yet the impact is still significant. A small founding group size and/or infrequent group mixing increases the variance in social behaviour between groups, and thus makes between-group competition a major driver of the evolutionary dynamics. After Fig. 3 in Powers et al. (2012).

within-group competition a larger driver of the evolutionary dynamics. In such cases, defection and ASP is favoured. Figure 2b shows us is that introducing ASP reduces the evolutionary contexts where cooperation is favoured even further.

Punishment is by no means required for cooperation. That cooperation is adaptive in a wide range of circumstances has been long understood. In fact, cooperation between replicators is necessary for the existence of any organism with more than one gene — that is, for all life (Dawkins, 1982). It is even easier to explain in multi-gene organisms (Hamilton, 1964). In our opinion, the most interesting implication of our simulations is that at an ultimate level, punishment can be used *either* to increase or decrease cooperative investment in public goods. In contexts where investment at the group level is unlikely to be beneficial (e.g. where public goods are likely to be confiscated by other groups before they are exploited), members of the group may receive better inclusive fitness benefits from more direct investment e.g. in offspring. This opens the door to the possibility that the proximate mechanisms that lead to punishment serve ultimately as a distributed mechanism for regulating the level of investment populations make to that most appropriate for an individual's socio-economic context.

3.4 Proximate Causes and Consequences of ASP

We now move from ultimate causes and simulations to proximate mechanisms and explorations of real human data. The first thing worth noting is that the terminology behind ASP and AP can be quite misleading (Sylwester et al., 2013b,a). Firstly, 'altruistic' punishment is not generally altruistic in intention. Proximately, costly punishment is frequently motivated by aggressive tendencies. Secondly, ASP is not always aimed at the top contributors, and cannot be ascribed entirely to revenge. ASP occurs even in the first round, before anyone has been punished. Sometimes ASP is aimed from the lowest contributor to the second-lowest contributor, in an apparent effort to make them produce more public goods while allowing the punisher to continue to free ride. When this occurs, ASP can actually be seen as an altruistic act, because the punisher pays a penalty, and the other members of the group (those who are not the punishee) benefit just as much as the punisher if the punishee increases their contribution. In fact, those who never punish (a sizeable minority) could also be seen as free-riders in cultures where punishment leads to an increase in the public good.

The fact that this terminology is misleading does not mean it should be abandoned. ASP and AP both have clear definitions (as given in our Introduction), and clear correlates with important measures of economic wellbeing (as demonstrated by Herrmann et al., 2008a). But we need to remember not to associate the obvious moralistic assessments with these terms, and more generally that socio-economic behaviour, dependencies and outcomes are highly complex. Further, as has often been noted, most contemporary human experimental subjects are University undergraduates, and universities have historically been most prominent in countries that are high in indices such a wealth, democracy and rule of law (Henrich et al., 2010). As Herrmann et al. (2008a) have shown, these

societies express relatively little ASP, and as a consequence antisocial punishment has been regarded as a marginal phenomenon, perhaps explicable simply as revenge taken by those punished altruistically (Fehr and Gächter, 2002).

One might expect that ASP would lead directly to reduced contributions just as AP leads to increases, but in fact victims' responses to ASP are much less directed than victims' responses to AP (Sylwester et al., 2013b). This indicates that the evolutionary 'strategy' associated with punishment expression may well included the punishment response. Without punishment, nearly eighty percent of individuals maintain from one round to the next their previous level of investment in the public good. However, among victims of ASP the number maintaining their previous-round's strategy falls to nearly forty percent, though the *direction* of change shows no clear pattern. In contrast, victims of AP reduce their probability of repeating their investment level to only twenty percent, and are much more likely to increase investment than to decrease it. These results are despite the fact the individuals in these experiments do not know who punished them, and therefore often cannot determine certainly whether their punishment was altruistic or antisocial[7]. However, notice that subjects with very low or very high prior levels of investment have only one direction to change in.

Given our hypothesis that punishment's expression may be determined by in group / out group assessment, we can mine a great deal of psychological literature for candidate proximate causes in the form of cues that trigger shifts in these assessments. Sylwester et al. (2013a) explain that we would expect AP to be less useful when applied to members of out groups, since it might prompt members of other groups to behave more cooperatively thus decreasing the punisher's own group's relative ranking and therefore (presumably, if there is group-level competition) resources. Conversely, we would expect ASP to be practiced less in contexts where the other group members are assessed as 'in group'. Lamba and Mace (2012) show empirical evidence supporting this idea. In extremely similar but discrete populations of a very small-scale minority culture in India (the Pahari Korwa), Lamba and Mace demonstrate lower levels of ASP in villages that contain a higher proportion of other cultures as well, compared to villages exclusively composed of Pahari Korwa. This may indicate that the presence in a village of a potential out group led subjects of a game played between members of a single culture to treat each other as in-group. But where the criteria for selection of experiment participation was not so clearly ethnic (due to only one ethnicity being present in the village), subjects were more likely to view each other more as potential competitors.

3.5 Individual Strategies: Variation in ASP Is Best Predicted by Proportions of Highly Cooperative Actors

Not every individual in a population will necessarily express the same behavioural strategy in the same immediate context. As explained earlier, we hypothesise

[7] Of course, one in four individuals give the least in their group so know any punishment is altruistic, and the same number contribute the most and know theirs is antisocial.

that at a population level, the ultimate explanation of variation in punishment strategies and their associated economic productivity is an optimising response to local political and economic conditions which can determine the expected outcome of a public good investment. Therefore, we should expect that whatever the proximate mechanism for selecting an investment (including punishment) strategy is, it should respond to evidence or experience indicating changes in this underlying context. Presumably, each individual responds to their own individual experience, though this may include the stories they hear from others and their upbringing. Their exact response may also be determined by other predispositions such as personality or self-assessed social ranking.

Notice therefore that we do not need to expect everyone in a population to express the same strategy at the same time. We only expect that first, the net result of combining these strategies in the proportions found in a population tracks the underlying context, and second that each of the strategies should be self-sustaining in the extent they are expressed within that context. MacLean et al. (2010) document how for even very simple organisms in a simple environment, it may be easiest to optimise exploitation of that environment by altering the number of individuals expressing pro- or anti-social behaviour, in the form of investing in or free riding on public goods.

If AP really did account for cooperative behaviour, we might expect its prevalence to correlate with economic performance, and that of free riding to be anti-correlated. In fact, we have found the reverse. In examining the dataset due to Herrmann et al. (2008a), we found both free-riding and AP to be fairly consistent across populations. What varies with regional economic performance as measured by GDP is the proportion of strong Cooperators in a society (defined below), and the propensity to anti-socially punish cooperators.

To investigate better correlates of decreased contributions we explored the hypothesis that subject pools might differ in the composition of cooperative types. For clarity (and after some experimentation), we focussed on distinguishing just two classes of extreme behavioural types from among the participants. Our classification was based on participants' behaviour in the very first round of the first public goods game they played, in cases where no punishment was allowed. All behavioural economics subjects must demonstrate full comprehension of a task in a test before they are allowed to participate in an economic game. The first move therefore signals better than anything else their expectations brought into the experiment — their interpretation of likely events as well as their own predispositions. After the first round, PGG subjects are known to demonstrate significant conformity bias (Carpenter, 2004; Bardsley and Sausgruber, 2005). Extreme contributors tend to move more towards the group average, though still maintaining a bias towards their initial action.

We classified those who invested their entire initial allocation to the group account as *Cooperators* (with a capital C). Those with who did not make any group investment at all we classified as exploitative *Free-riders*. The rest (the vast majority of participants) we did not classify. We reasoned that if a person devotes their entire allocation to the group welfare, full cooperation is likely their

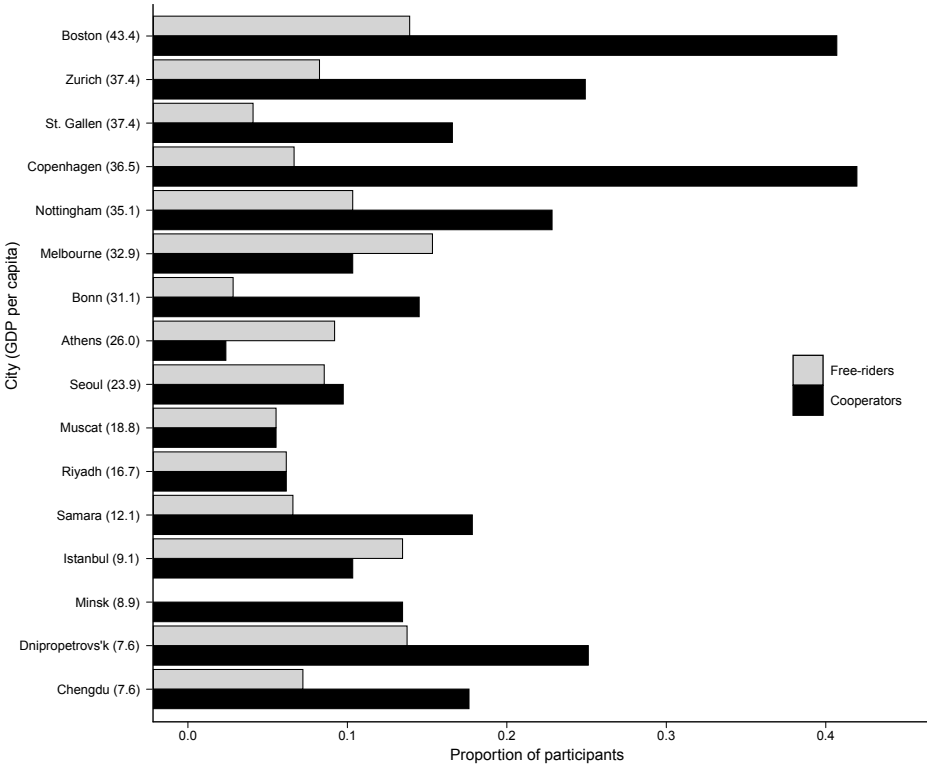


Fig. 3: Proportion of participants that contributed all (Cooperators) or nothing (Free riders) by city from the Herrmann et al. (2008a).

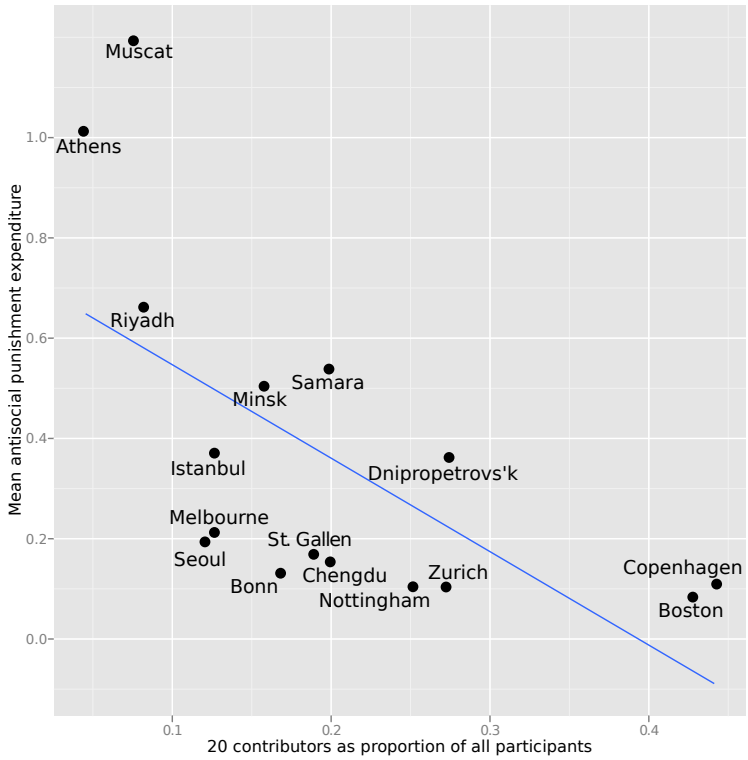


Fig. 4: Subject pools plotted by mean amount of ASP (y axis) and the proportion of subjects who contributed all of their available resources (20 tokens) in the first round.

default behaviour when interacting with strangers. Analogously, we assumed that people who do not make any effort to support their new group have a tendency to behave in an exploitative fashion, or at least not to trust others to cooperate.

We found that the variation across subject pools in the proportion of Co-operators is much greater than the variation in the proportion of Free-riders (see Figure 3), Levene's test $= 6.71$, $p = 0.01$; $MFREE - RIDERS = 0.10$, $SD = 0.05$, $MCOOPERATORS = 0.20$, $SD = 0.11$. We then ran correlations, to determine whether there is a link between the proportion of cooperative types in a subject pool and the mean expenditure on ASP. The correlation between AP and the proportion of Cooperators ($r = 0.35, p > 0.05$) was not significant. Neither was the correlation between AP and Free-riders ($r = -0.18, p > 0.05$), nor between the proportion of Free-riders and ASP ($r = -0.20, p > 0.05$). In contrast, we found a strong anticorrelation between the proportion of Cooperators and ASP ($r = -0.62, p < 0.01$, Figure 4).

This means contrary to expectation that the variation between cultures may be primarily the difference between the probability of individuals playing an

optimistic, Cooperative strategy. Such behaviour may *inhibit* expression of ASP even in regions / socio-economic contexts where we had hypothesised unexpected generosity served as a trigger — where it was likely to be viewed as a competitive or dominance-seeking act. Perhaps extreme cooperation signals in-group affiliation. However, anticorrelation does not allow us to infer causation. It may be that expecting antisocial punishment inhibits reckless tendencies for Cooperation. Our findings do however suggest more environmentally-determined plasticity in the proportion of individuals with cooperative, rather than exploitative, predispositions. A multiple regression shows that a number of socio-economic factors predict the proportion of Cooperators but not Free-riders. Our analysis is the first to demonstrate that the distributions of extreme cooperative, but not uncooperative, tendencies differ across human populations.

4 Summary and Implications

In the previous section we documented our contributions to the behavioural anthropology of human economic decision making, many of which derive from our taking an evolutionary approach and perspective. The assumption of this work is the standard one made in biology: that the seemingly bizarre behaviour of ASP must be a part of a behavioural strategy that is generally advantageous — or at least not disadvantageous — to people living in some cultural contexts, presumably the ones in which it is found. To briefly summarise some of our findings:

- ASP has a disruptive more than a reliably 'down-regulating' influence on cooperation. It does not reduce cooperation as reliably as AP increases it, but it does tend to alter investment behaviour, though again AP is even more likely to result in changed behaviour.
- Down regulating cooperation might make sense for an individual if that individual's wellbeing is determined more by local competition (e.g. who is most dominant in a household, village or business) than by global competition (e.g. which household, village or business does best.)
- ASP seems to be expressed more frequently in contexts where group members do not by default expect other group members to be members of their in group. With respect to the previous point, this implies that there is always *some* cohort of trusted individuals, the question may be how large that cohort is by default. In Northern Europe (and Boston, the only US city surveyed here), the in group seems to encompass group sizes at least as large as a single university, while in Greece, Turkey, the Middle East and the former Soviet Union it does not.
- Whatever the *default* level of in-group assessment is, some manipulations might alter this. The only ones we could explore without performing human subject experiments was the natural experiment of seeing how subjects respond to having someone in their group who contributes *all* of their resources to the public good, and of having someone in the group who contributes

none. Interestingly, we have learned here that having super-defectors in the group has no effect, but having super-cooperators in the group is correlated with reduced level of ASP. This might mean that people inclined to ASP are impressed by such a clear expression of in-group assessment, and have some tendency to believe and adopt it. This hypothesis needs to be checked.

4.1 Applications to Policy

This last point — that manipulations which increase the probability of in-group assessment might also decrease levels of ASP and increase economic viability overall — is one of obvious potential, should experimentation bear it out. The one potential manipulation for which we have any data though may be difficult to replicate. Even if having group members that make extremely altruistic contributions does inhibit competitive tendencies rather than just covarying with such inhibition, the experimental context is highly unusual because of its transparency. All subjects know they have equal access to information and equal power under the authority of the experimenter. In a more realistic context, showing total economic commitment or some other signal of in-group affiliation may be difficult to control, particularly by outsiders. There may however be other team-building exercises that would have at least local efficacy in facilitating negotiations.

Many people likely to read this chapter can recognise and identify with the in-group assessment apparently made by subjects from Boston (Harvard) and the Northern European universities tested. Knowing someone else has chosen the same college or university as we have, particularly in the same or similar year, does indicate a likely similarity. An undergraduate degree is a significant investment — even where tuition is free, a degree requires 3–5 years of a person's life. For many of us, making similar investments at this scale is enough to incline us towards in-group trust, but then we live in societies with a high Rule of Law (cf. Section 2). Understanding the social experience of those who cannot make this assumption about their colleagues requires effort for those who can. However, almost anyone will have had *some* experience of being in a situation where we were not sure everyone in the room was interested in collaborating for our mutual common good — where we have felt in danger of exploitation. Realising that in some cultures that feeling appears to extend even to the prestigious university campuses that Herrmann et al. (2008a) chose to study[8]. This might indicate that it could be unexpectedly challenging to achieve trust and therefore high levels of economic cooperation in other professional contexts as well as a university.

We must remember that in every society studied, ASP was practiced by some participants, but similarly in no society was it practiced by all. It may be that further experiments will identify in advance personality indicators for predisposition to ASP (e.g. Czibor and Bereczkei, 2012). On the other hand,

[8] Because the initial studies were conducted at ETH, it was considered essential that representatives from other cultures were also drawn from top universities to increase comparability.

these may not exist. ASP may respond primarily to a combination of present and cultural context, combined with an element of stochasticity. However, even if we could determine who practices ASP, we have no idea of what the broader impact for a society would be if these individuals were excluded from positions of power or negotiation. As we mentioned, in some circumstances reducing group size or down-regulating public investment may make economic sense, thus those able to recognise this may be important members of a society or organisation.

We also do not know for sure that decreasing ASP and/or increasing co-operation would increase GDP. The causality could well be reversed — where individuals are affluent they can take more risks about in-group inclusiveness. It seems likely though to be a situation of mutual feedback, and that if honest, transparent signals of mutuality of interest can be established, higher levels of both cooperation and economic performance could be established.

4.2 Conclusion

> To have received from one, to whom we think ourselves equal, greater benefits than there is hope to requite, disposeth to counterfeit love, but really secret hatred, and puts a man into the estate of a desperate debtor that, in declining the sight of his creditor, tacitly wishes him there where he might never see him more. For benefits oblige; and obligation is thraldom; and unrequitable obligation, perpetual thraldom; which is to one's equal, hateful. But to have received benefits from one whom we acknowledge for superior inclines to love; because the obligation is no new depression: and cheerful acceptation (which men call gratitude) is such an honour done to the obliger as is taken generally for retribution... — Hobbes (1651)

Our work has shown that Hobbes was amazingly prescient concerning the creation of public goods given that he wrote in the seventeenth century, but not entirely right. Our research indicates that anti-social punishment may indeed occur in contexts where other participants are not mutually-acknowledged members of trusted group, and a gift from an anonymous peer may be met with suspicion or loathing. However, we have also seen that generosity may in absence of other information be taken as an indication that in fact trust is merited, and generosity should be accepted.

We have found that costly punishment is best understood as having impact not only on global economics but also on individual competition, and that the apparently-maladaptive behaviour of anti-socially punishing those more generous than ourselves may even in some contexts be a sensible response. When an actor's own wellbeing is (or at least appears to be) most determined by their relative dominance to their local neighbours, rather than to how well the neighbourhood performs as a whole, then it may be worth sacrificing immediate opportunities if longer-term benefits e.g. in terms of in-group status result. For organisations that *are* more concerned about global than local good, the best course of action is probably to first promote the likelihood that the benefits of

public goods are shared by those who invest in them, and second to promote transparency, so that all parties involved in investment decisions can be assured that their interests are protected.

Throughout this chapter we have taken the perspective that the failure to find communal economic optima is fundamentally negative, since it means resources are wasted in conflict and all parties have less access to wealth and its associated wellbeing. Assuming this, the most directly applicable avenue for future research might be to discover how easily or quickly the social characteristics leading to this failure can be altered. Measures available could be either either cognitive, such as increased transparency or reliability in the distribution of economic resources, or emotional, such as team building or other stage setting for triggering a state of emotional inclusiveness. If such measures work, a societies' citizens and / or leaders could be trained to recognise, exploit or promote contexts where mutually advantageous outcomes were possible. However, it may be that for some societies such interventions would be impossible, impractical or unethical. Even in such cases, we could at least hope that the outcome of research in this area would still be beneficial. It would help us to at least identify, characterise and possibly come to understand cultures with such differences. This might be useful for selecting strategies in cross-party negotiations, or in choosing between economic policy options or approaches to development.

Acknowledgements

We would like to thank Benedikt Herrmann for his advice and help with theory building, the literature, and his assistance in understanding his own data set. Thanks also to Simon Gächter for meetings and occasional email assistance, and Daniel Taylor for many conversations and useful analysis. Thanks to Will Lowe for his help with data, statistics, software and analysis, and to Gideon Gluckman for support in writing. From October 2010–September 2011, much of this effort was supported by the US Air Force Office of Scientific Research, Air Force Material Command, USAF, under grant number FA8655-10-1-3050. We would also like to thank the Department of Computer Science and the University of Bath for further financial support.

Bibliography

Abbink, K. and Sadrieh, A. (2009). The pleasure of being nasty. *Economics Letters*, 105(3):306–308.

Barclay, P. (2006). Reputational benefits for altruistic punishment. *Evolution and Human Behavior*, 27(FIXME):325–344.

Bardsley, N. and Sausgruber, R. (2005). Conformity and reciprocity in public good provision. *Journal of Economic Psychology*, 26(5):664–681.

Boehm, C. (1999). *Hierarchy in the Forest: The evolution of egalitarian behavior*. Harvard University Press.

Bryson, J. J. (2009). Representations underlying social learning and cultural evolution. *Interaction Studies*, 10(1):77–100.

Bryson, J. J., Ando, Y., and Lehmann, H. (2007). Agent-based models as scientific methodology: A case study analysing primate social behaviour. *Philosophical Transactions of the Royal Society, B — Biology*, 362(1485):1685–1698.

Bryson, J. J., Ando, Y., and Lehmann, H. (2012). Agent-based models as scientific methodology: A case study analyzing the DomWorld theory of primate social structure and female dominance. In Seth, A. K., Prescott, T. J., and Bryson, J. J., editors, *Modelling Natural Action Selection*, pages 427–453. Cambridge University Press.

Carpenter, J. P. (2004). When in rome: conformity and the provision of public goods. *The Journal of Socio-Economics*, 33(4):395–408.

Clutton-Brock, T. H. and Parker, G. A. (1995). Punishment in animal societies. *Nature*, 373(6511):209–216.

Czibor, A. and Bereczkei, T. (2012). Machiavellian people's success results from monitoring their partners. *Personality and Individual Differences*, 53(3):202–206.

Dawkins, R. (1982). *The Extended Phenotype: The Gene as the Unit of Selection*. W.H. Freeman & Company.

Fehr, E. and Gächter, S. (2000). Cooperation and punishment in public goods experiments. *The American Economic Review*, 90(4):980–994.

Fehr, E. and Gächter, S. (2002). Altruistic punishment in humans. *Nature*, 415(6868):137–140.

Gintis, H., Bowles, S., Boyd, R., and Fehr, E. (2003). Explaining altruistic behavior in humans. *Evolution and Human Behavior*, 24(3):153–172.

Hamilton, W. D. (1964). The genetical evolution of social behaviour. *Journal of Theoretical Biology*, 7:1–52.

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., and McElreath, R. (2001). Cooperation, reciprocity and punishment in fifteen small-scale societies. *American Economic Review*, 91(2):73–78.

Henrich, J., Heine, S. J., and Norenzayan, A. (2010). Most people are not WEIRD. *Nature*, 466(7302):29.

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

Herrmann, B., Thöni, C., and Gächter, S. (2008b). Supporting online material for antisocial punishment across societies. *Science*, 319(5868). http://www.sciencemag.org/cgi/content/full/319/5868/1362/DC1.

Hobbes, T. (1651). *Leviathan*. Andrew Crooke, London.

Jensen, K. (2010). Punishment and spite, the dark side of cooperation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365(1553):2635–2650.

Kaufmann, D., Kraay, A., and Mastruzzi, M. (2004). Governance matters III: Governance indicators for 1996, 1998, 2000, and 2002. *The World Bank Economic Review*, 18(2):253–287.

Kokko, H. and López-Sepulcre, A. (2007). The ecogenetic link between demography and evolution: can we bridge the gap between theory and data? *Ecology Letters*, 10(9):773–782.

Laland, K. N., Sterelny, K., Odling-Smee, J., Hoppitt, W., and Uller, T. (2011). Cause and effect in biology revisited: Is Mayrs proximate-ultimate dichotomy still useful? *Science*, 334(6062):1512–1516.

Lamba, S. and Mace, R. (2012). The evolution of fairness: explaining variation in bargaining behaviour. *Proceedings of the Royal Society B: Biological Sciences*.

Ledyard, J. O. (1995). Public goods: A survey of experimental research. In Kagel, J. H. and Roth, A. E., editors, *Handbook of Experimental Economics*, pages 111–194. Princeton University Press, Princeton, NJ.

MacLean, R. C., Fuentes-Hernandez, A., Greig, D., Hurst, L. D., and Gudelj, I. (2010). A mixture of "cheats" and "co-operators" can enable maximal group benefit. *PLoS Biol*, 8(9):e1000486.

Martin Clarke, D. E., editor (1923). *Hávamál*. Cambridge University Press.

Mauss, M. (1967). *The Gift: The Form and Reason for Exchange in Archaic Societies*. WW Norton, New York. Ian Cunnison, translator.

Mayr, E. (1961). Cause and effect in biology: Kinds of causes, predictability, and teleology are viewed by a practicing biologist. *Science*, 134(3489):1501–1506.

Powers, S. T., Penn, A. S., and Watson, R. A. (2011). The concurrent evolution of cooperation and the population structures that support it. *Evolution*, 65(6):1527–1543.

Powers, S. T., Taylor, D. J., and Bryson, J. J. (2012). Punishment can promote defection in group-structured populations. *Journal of Theoretical Biology*, 311:107–116.

Preuschoft, S. and van Schaik, C. P. (2000). Dominance and communication: Conflict management in various social settings. In Aureli, F. and de Waal, F. B. M., editors, *Natural Conflict Resolution*, chapter 6, pages 77–105. University of California Press.

Rand, D. G., Armao IV, J. J., Nakamaru, M., and Ohtsuki, H. (2010). Antisocial punishment can prevent the co-evolution of punishment and cooperation. *Journal of Theoretical Biology*, 265(4):624–632.

Rand, D. G. and Nowak, M. A. (2011). The evolution of antisocial punishment in optional public goods games. *Nature Communications*, 2:434.

Rohwer, Y. (2007). Hierarchy maintenance, coalition formation, and the origins of altruistic punishment. *Philosophy of Science*, 74(5):802–812.

Sylwester, K., Herrmann, B., and Bryson, J. J. (2013a). *Homo homini lupus?* Explaining antisocial punishment. *Journal of Neuroscience, Psychology, and Economics*. accepted for publication.

Sylwester, K., Mitchell, J., and Bryson, J. J. (2013b). Punishment as aggression: Uses and consequences of costly punishment across populations. submitted.

Sylwester, K. and Roberts, G. (2010). Cooperators benefit through reputation-based partner choice in economic games. *Biology Letters*, 6(5):659–662.

Taylor, D. J., Richards, M., and Bryson, J. J. (2013). Does reciprocation explain cooperation in large groups? in prep.

Thierry, B. (2005). Integrating proximate and ultimate causation: Just one more go! *Current Science*, 89(7):1180–1183.

Čače, I. and Bryson, J. J. (2007). Agent based modelling of communication costs: Why information can be free. In Lyon, C., Nehaniv, C. L., and Cangelosi, A., editors, *Emergence and Evolution of Linguistic Communication*, pages 305–322. Springer, London.

Walker, A. and Stringer, C. (2010). The first four million years of human evolution. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365(1556):3265–3266.

West, S. A., El Mouden, C., and Gardner, A. (2011). Sixteen common misconceptions about the evolution of cooperation in humans. *Evolution and Human Behavior*. in press.

West, S. A., Griffin, A. S., and Gardner, A. (2007). Evolutionary explanations for cooperation. *Current Biology*, 17:R661–R672,.

Whitehouse, H., Kahn, K., Hochberg, M. E., and Bryson, J. J. (2012). The role for simulations in theory construction for the social sciences: Case studies concerning Divergent Modes of Religiosity. *Religion, Brain & Behavior*, 2(3):182–224.