Understanding and Addressing Cultural Variation in Costly Antisocial Punishment

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

[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 human uniqueness relative to other species. There is no question that our level of investment in culture is unique, however in fact humans will sometimes punish those who contribute to the common good. This behaviour — antisocial punishment — is negatively correlated with GDP, and as such may be seen as a hinderance to overall wellbeing that needs to be understood and addressed. In this chapter we exploit existing data showing cultural variation in propensity to punish to ask how such sanctioning, whether of those who give much or little, affects the individuals who conduct it. We hypothesise that costly punishment is in fact a mechanism for regulating investment between different levels of society — whether current focus should be on the nation, village, family or the self. We suggest that people are less likely to antisocially punish those they consider to be "in group", and suggest that this measure of identity may vary with socio-economic-political context to regulate investment. 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 well being.

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

1 Introduction

The variety of human cultures is one of the joys of contemporary human life. However, our respect and appreciation for diversity does not stop us from observing 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 find ways to optimise economic collaboration when thrown into a group together. In this chapter we review our progress so far, and examine policy implications of this, particularly for organisations interested in aiding development or rebuilding communities in areas experiencing conflict.

The behaviour we are studying is called *anti-social punishment* (ASP). This 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 *more generous* than the punisher. That is, the punisher is paying a cost to damage the interests of an individual who has given the punisher economic advantage. Cultural variation in this behaviour was first reported by Herrmann et al. (2008), 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 at least that the behaviour measured in the laboratory may have fundamental impact on the economic well-being of a nation, though of course the reverse could also be true. Further, the variation between cultures is not arbitrary, but rather seems to be clustered by global region. Thus Boston, several cities in Northern Europe, the Far East, and Melbourne show high levels of profitable and altruistic economic collaboration, while Athens, Istanbul, regions of the Middle East and of the former Soviet Union show relatively low levels, and higher levels of expression of ASP.

This chapter begins with a review the 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. The bottom line is that 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 must be social status, and is probably awarded to those who punish regardless of whether they do so altruistically (punishing those who give less than themselves to the group) or antisocially. If we include this assumption, then we *are* able to account for variation in ASP. Variation would be expected to track the extent to which ones well-being depends more on ones relative status with one's own group, or the relative status of one's group overall.

After reviewing these finds, we discuss policy implications for our work. Clearly the global economy benefits when local economies produce mutual benefit rather than destruction, but given our new understanding of why pathways to mutual benefit may not always be recognised, what can governments and

outside organisations do? We make a number of suggestions, then close with our conclusions.

2 Scientific Background: Costly Punishment

Herrmann et al. (2008) showed that in some human subject pools (e.g. university undergraduates in the Boston, Melbourne, Chengdu and Zurich) 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 the society will pay a penalty in order to further penalize others who are being more generous than themselves. This is despite the fact that this generosity is benefiting all group members *other than* the benefactor, including the punishers. Such punishment of altruism is called *antisocial punishment* (ASP).

Herrmann et al. sought correlates for the prevalence of ASP in a culture, and found that the two strongest are that it inversely correlates to 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 show 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?

As we discuss below in Section 3.3, the disruptive findings of Herrmann et al. created a theory vacuum. We therefore must build new theories — preferably literally, as models that can be tested for viability and against the data. But first we describe the data of in more detail.

2.1 The Data: How Altruism and Antisociality are Measured

All human subject data for this research was collected using a paradigm from a relatively new branch of economics, generally called either *experimental economics* or *behavioural economics*. This is somewhat like classic experimental psychology, except a few special protocols are followed:

- The performance of all individuals must be rewarded by a sufficiently-significant financial reward that there can be no doubt that financial motivation is present.
- There can be absolutely no deception of subjects. The standard psychological 'tricks' of telling subjects one thing is the task while measuring something else are not allowed. In fact, if there has been any deception committed against subjects in the subject pool (e.g. any history of non-payment or unexpected payment) the entire university is blacklisted from the economics literature.

– All subjects must demonstrate understanding of the complete consequences of their own and the other team members' actions. This is done by means of a test after training. Subjects that cannot pass the test are excluded from the experiment. Thus we know that subjects know exactly what they are doing, and what benefits or sacrifices they are making for themselves depending on their actions.

In cross-cultural experimental economics research, the players play for tokens, for which they know the value in local currency. The reason they play for tokens is that it is easier to reason about how to divide 20 tokens than (for example) $7.82. Thus the motivation is provided by local currency at levels set by local standards of living, but the numeric reasoning is supported to be similar across cultures.

The standard type of behavioural economics experiment for assessing costly punishment is called the Public Goods Game (PGG). In the basic form of this game there is no punishment. PGG represents a social dilemma because individual interests are in conflict with the groups interests. 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[3]. This anonymity is maintained to ensure group members do not act out of fear or expectation of retribution after the game. In a single round of PGG, each member is allocated 20 tokens, and then individuals are allowed to contribute any portion of their allocation to the public pool. Allocations to the public pool are multiplied by the experimenter then divided equally between all group members. However, the multiplication is never so great that an individual receives as much money back from their own investment as they paid. Thus individuals who do not contribute anything or contribute less than others gain a financial advantage, at least for that round. PGG may be played as a single round, but for the results described here they are played in ten rounds, all with the same group.

In the punishment condition, after a round of PGG individuals can anonymously punish others. This punishment is not based on identity, but only on the other's contribution to the public pool in the most recent round. Importantly, subjects never learn any information about who punishes them, only about what others have contributed to the pool. In the studies described here, the cost/effect ratio is that for every token a punisher pays, the punishee loses three tokens[4]. When an individual punishes someone who has contributed less than they have, this punishment is 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 more cooperative contributions. On the other hand, if punishers punish those who contribute more than they do, this is called *anti-social*. Herrmann et al. (2008) were the first to document societies with large amounts of ASP, and

[3] 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.

[4] Many other ratios have been tried by other experimenters, these result quantitative but not qualitative shifts in behaviour. See (Sylwester et al., 2011) for a more complete review.

showed that this could in some cases completely counter the expected benefits of cooperation. In the Swiss contexts where these experiments were first run, the punishment condition of the PGG reliably resulted in a better economic outcome for the subjects, but 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 Earlier Interpretations of Punishment Results

To understand the full literature and history of work in costly punishment, it must first be understood that one of the 'holy grails' 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, given that very human-like species existed and used primitive tools for millions of years. Urbanisation, agriculture, writing and doctrinal religion (religions shared outside 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 emphasised the human propensity to cooperate as a possible explanation for the presence of culture (e.g. Gintis et al., 2003; Henrich et al., 2001). However, this only presents a chicken-and-eggs problem — is this an explanation or a redescription? What accounts for this level of cooperation? After the early PGG results (e.g. Fehr and Gächter, 2000; Fehr and Gächter, 2002), altruistic punishment was originally regarded as a possible explanation for cooperation. Here too though the reasoning seemed cyclic, as punishment was a form of altruism itself. Contrary to its reputation, altruistic behaviour neither difficult to evolve nor uniquely human (Čače and Bryson, 2007; West et al., 2007; Bryson, 2009). The Herrmann et al. (2008) indeed indicate that punishment is part of a much more complex system of social regulation rather than a simple explanation for human culture. More recently, the phenomenon of ASP has 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). Swings of moral assessment and defensiveness need to be guarded against if we are to understand what underlies these phenomenon. The research presented here attempts an objective perspective, by approaching the problem as rooted in the ultimate explanation known for biological phenomena: natural selection. In the next section, we review the sorts of explanations natural selection can provide for behaviour.

2.3 Proximate and Ultimate Explanations

Although like much in evolutionary biology the exact concepts and terms presented here are still subject to debate and refinement, there is broad agreement

on their general meaning. Any behaviour we see in nature is generally expected to have a number of different types of cause. *Ultimate causes* concern why the behaviour is present in a population as whole — what role does it serve in the evolutionary struggle? Note that 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, that an extant trait exists because it has historically provided more advantage than disadvantage to those who hold it relative to those that do not is at least a common first guess in evolutionary approaches. A *Proximate cause* is the 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.

Note also that a useful proximate mechanism may itself become an ultimate explanation for some other trait. For example, the utility of running from some noises but not others may result in selective pressure for being able to better discriminate between these two sorts of noise.

3 Building an Understanding of Anti-Social Investment

In this section we introduce our findings concerning an explanation for anti-social punishment. Our current hypothesis is that all punishment is an aggressive act, which some proportion of any population is motivated to perform. One reward for aggression is increased social status, not only relative to the target of the aggressive act but also to bystanders who, on witnessing aggression, will associate an increased cost with confronting the aggressor. However, in contexts where cooperation is likely to produce value, more of the population will 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.

This assessment we believe correlates to how locally an individual will choose to invest available resources such as effort. One way to think about local versus global (in the biological sense) competition is that local competition occurs *within* groups — e.g. who in my family gets the biggest piece of pie? In contrast, global competition occurs *between* groups — e.g. 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 nation against another.

Another relavant concept is the *zero-sum assumption*. If there is only one pie of a fixed size, then the size of my piece is the only thing that affects my personal

well being (zero-sum), thus for me to gain more someone else has to lose. This can lead to competition being the only viable strategy. If there is a way to create more or bigger pies, then it could very well be worth collaborating to do so. In either case, the pies are a public good, what varies is not only the strategy for exploiting them, but also the return on investment for contributing to baking them. Competition like collaboration is always costly; thus where the expected gains from collaboration outweigh those from competition for the same amount of risk and effort, collaboration is a viable strategy.

Recall that all of the data that our experiments come from anonymous individuals on university campuses. Therefore what we are supposing varies culturally is

1. *proximately*: the default assumption about strangers in the context of an economic experiment at a prestigious university, and
2. *ultimately*: the underlying socio-political-economic context which reward or penalise sets of assumptions, and thus explain regional variation in distributions of these assumptions, presumably due to varied individual experiences and / or familiar narratives.

Where we describe human data results, these are derived from the original Herrmann et al. (2008) data set with additional analysis, mostly by Sylwester and Mitchell. We also describe results derived from agent-based modelling (ABM). ABM is a process of describing a model so thoroughly that its consequences can be determined through sampling by simulating the model on a computer. All modelling (whether simulation, verbal or mathematical) provides information about theories — the output of simulations however also serves as predictions, which can be compared to the real world. Computer modelling also allows us to check for internal coherence of our theories, since inconsistant theories are impossible to build and run as programmes. Whitehouse et al. (2012) give further general-purpose information on ABM in the social sciences. Here, most modelling has been performed by Powers and Taylor.

3.1 The Terminology Behind ASP and AP Is Misleading

The first thing evident from examining the data is that the terminology behind ASP and AP are quite misleading Sylwester et al. (2013). *Altruistic* punishment is not generally altruistic in intention. Proximately, all costly punishment seems frequently motivated by aggressive tendencies. Indeed, punishment may also have consequences in establishing social dominance, whether or not that is its intended purpose. 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. Speaking strictly to the biological definition of altruism, 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.

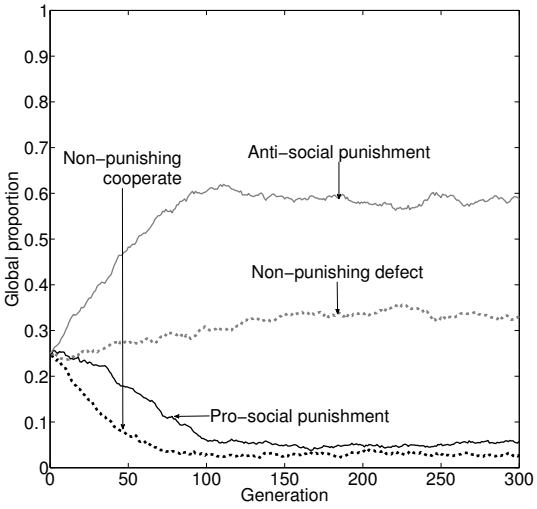
The fact that this terminology is misleading does not mean it should be abandoned. ASP and AP both have clear definitions and clear correlates with important measures of economic well being. But we need to remember that these terms cannot be used for obvious moralistic assessment and that socio-economic behaviour and dependencies are highly complex.

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

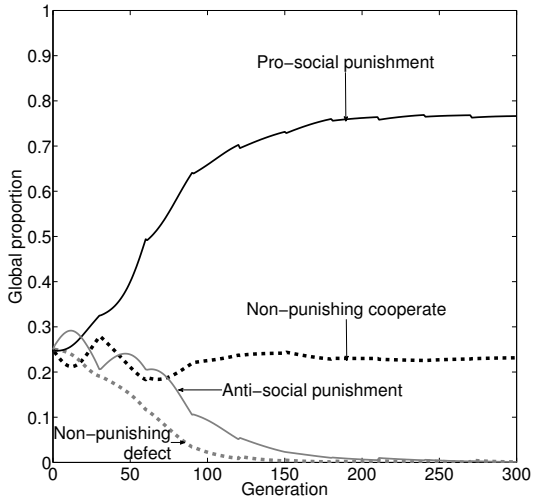
The results of this section were based on multi-level evolutionary agent-based models. These allow us to vary the relative importance of within-group competition and between-group competition. The set of models extend from the the work of Powers et al. (2011). Here, within-group competition is increased by *increasing* the group size, because the reward from cooperation is assumed to be fixed, so in a larger group the benefits of cooperation are reduced. As with many multi-level models, between-group competition is increased by decreasing the probability that individual agents find themselves in new groups, that is, by reducing the frequency with which groups are reformed (Szathmáry, 2011).

The particular models shown here are identical to those used by Powers et al. (2012). These evolutionary models of competition between pro- and anti-social punishment operate in a group-structured population. A linear public goods game with punishment is played within groups once per generation. The payoffs from this game determines the fitness of individuals, such that individuals with a high absolute payoff produce more offspring. Groups remain together for a fixed 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 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 of 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.

As explained in Section 3.1, ASP is definitionally costly and, relative to other group members, altruistic, since they too benefit by the loss of relative fitness of the punishee, yet pay no costs themselves. 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 a slight benefit to the punisher in order for ASP to ever be adaptive. 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.



(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.

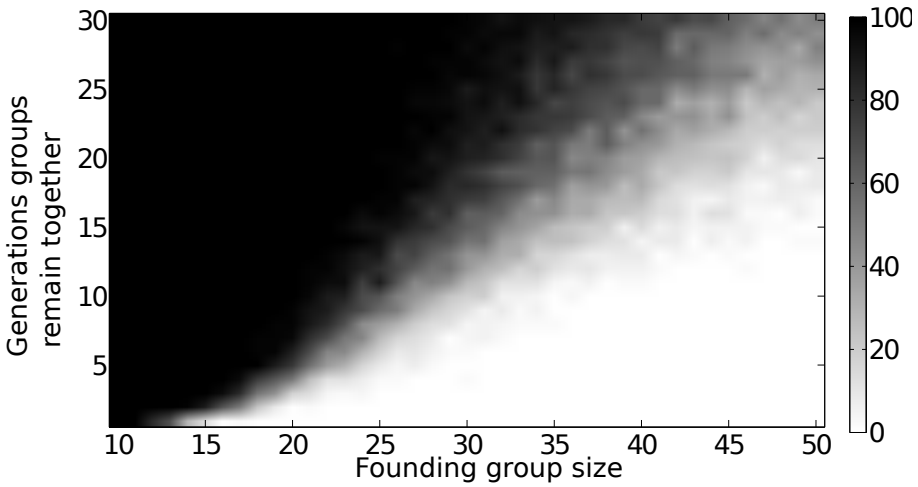
Our best guess at the moment is that punishment is used to signal or even generate dominance within a group. The benefits of social dominance over the lifetime of the punisher may more than compensate for the immediate cost of the punishment act (West et al., 2011). Indeed, dominance is often seen as 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). Dominance factors into the risk assessment of individuals entering social interactions. Publicly displaying agression in the form of punishment would increase an individual's reputation for being aggressive and the associated expected cost of entering into a dispute with them. Thus, both pro- or anti-social punishment 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).

In the case where punishment *does* result in intrinsic benefit (assumed to be associated with social dominance) then there is still an impact of local versus global competition. Where groups competer with each other — that is (in this theoretical context), where they persist long enough to exploit 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).

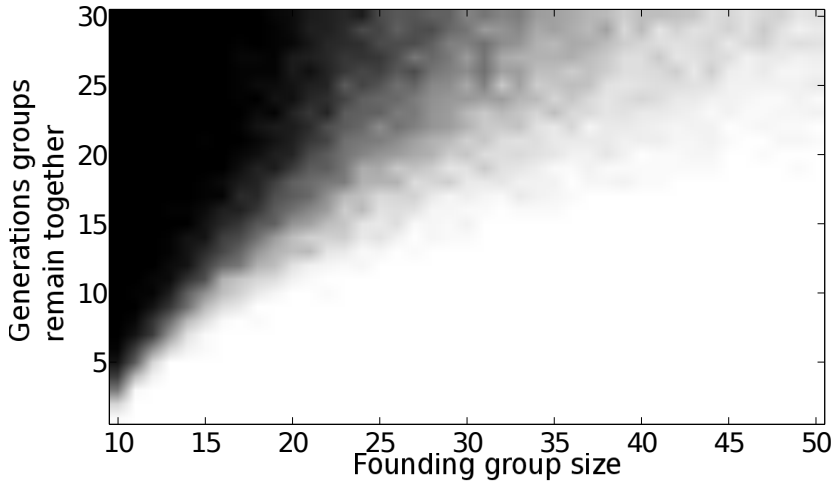
3.3 Punishment Alone Cannot Account for Human Sociality

A disruptive finding that follows from the above is that the fact that some individuals play ASP strategies completely undermines the theory described in Section 2.2 — that punishment explains humanity's exceptional levels of cooperation. Herrmann et al. (2008) demonstrated that populations exist in which the introduction of punishment actually reduces the level of public goods investment. 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, even in theory , once ASP is taken into account punishment in itself 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. Where group-size is relatively small (and thus, in our model, the individual share of the public good is relatively large) and relatively stable (there are many generations between dispersal and group reformation) cooperative strategies reliably evolve, otherwise they do not. Another way to describe this is that such conditions decrease the variance in social behaviour within groups, while increasing the variance between groups. As a result, cooperation and pro-social punishment is more likely to benefit. From a psychological point of view, this is similar to a situation in which there is a strong within-group identity, and hence a strong in-group / out-group distinction. Conversely, a large founding group size and/or frequent group-mixing



(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).

increases within-group variation in social behaviour, and hence makes within-group competition a larger driver of the evolutionary dynamics. In such cases, defection and anti-social punishment is favoured. This parallels a situation in which in-group identity is weak.

What Figure 2b shows us is that introducing ASP reduces the evolutionary contexts where cooperation is favoured even further. Notice that by no means is punishment required for cooperation, that cooperation is adaptive in a wide range of circumstances has been long understood. In fact, it is necessary for the existence of multi-gene organisms — that is, for all organisms. In our opinion, the fundamental result is that at an ultimate level, punishment can be used *either* to increase or decrease cooperation, that is it can be seen as a distributed mechanism to regulate the level of investment societies make to one appropriate to their socio-economic context.

There can be a context where too much is invested in public goods. To take one familiar to anyone who travels by air, it really is essential to put your own oxygen mask on (invest as an individual) before you can be competent to invest in others. Every population studied has shown positive levels of altruistic cooperation. What varies is how much is expressed, and thus the question is what indicators control this level of expression in human populations.

3.4 Proximate Causes and Consequences of ASP

Simulations are most useful for exploring ultimate explanations for a phenomenon. Proximate explanations should be relatively easy to explore experimentally, giving us better access to real data. However, most 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. (2008) 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 victim's response to ASP was much less directed than victim response to AP (Sylwester et al., 2013). This indicates that part of the "strategy" associated with punishment expression is actually 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, but when the subject of ASP that number falls to nearly forty percent, though the direction of change is essentially undetermined. Victims of AP however reduce their probability of repeating their investment level to only twenty percent, and are much more likely to increase investment than to decrease it. This is despite the fact the individuals in these experiments do not know who punished them, and whether it was altruistic or antisocial. However, notice that a subject with a very low prior level of investment has no direction to change in but up.

Now that we are hypothesising that punishment's expression may be determined by in group / out group assessment, we can mine a great deal of psychological literature for proximate causes in the form of cues that trigger shifts in these assessments. Sylwester et al. (2011) 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 contained a higher proportion of other cultures as well, compared to villages that exclusively composed of Pahari Korwa. This may indicate that the presence of a potential out group made a game played between members of a single culture be treated as in-group games, but where obvious out-groups were non existant in the broader population, subjects viewed their co-culture-members as potential competitors rather than collaborators.

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 strategy. We expect that the ultimate explanation of variation in punishment strategies and their associated economic productivity is an optimising response to local economic conditions, and to other societal conditions that determine who can expect to benefit from public good investment. Therefore, we should expect the proximate selection of strategy to be able to track changes in this. Presumably, each individual responds to their own individual experience (including though the stories they hear from others), though their exact response is also determined by their upbringing and other predispositions.

Notice therefore that we do not need to expect everyone in a population to express the same strategy at the same time. We only need to expect that the net result of combining these strategies in the proportions found in a population tracks the underlying context, and 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 a particular behaviour.

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. (2008), 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, and the propensity to anti-socially punish cooperators.

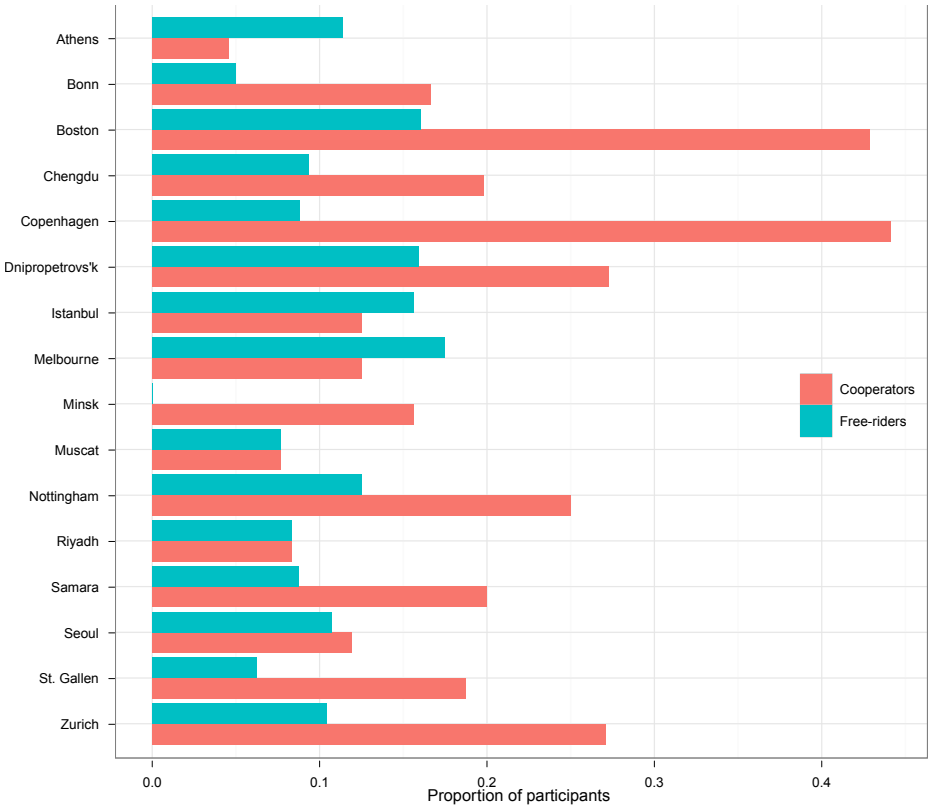


Fig. 3: Proportion of participants that contributed all (Cooperators) or nothing (Free riders) by city from the Herrmann et al. (2008).

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 their interpretation of likely events as well as their own predispositions. After the first round, we tend to see a good deal of conformity bias — extreme contributors tend to move more towards the group average, but still maintain 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), while those with who did not make any group investment at all, as exploitative *Free-riders*. The rest (the vast majority of participants) we did not classify. We reasoned that if a person devotes their whole allocation to the group welfare, full cooperation is likely their 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 Cooperators 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 actually *inhibit* expression of ASP rather than trigger it as a competitive or dominance-seeking act, perhaps by signalling 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 environmental 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.

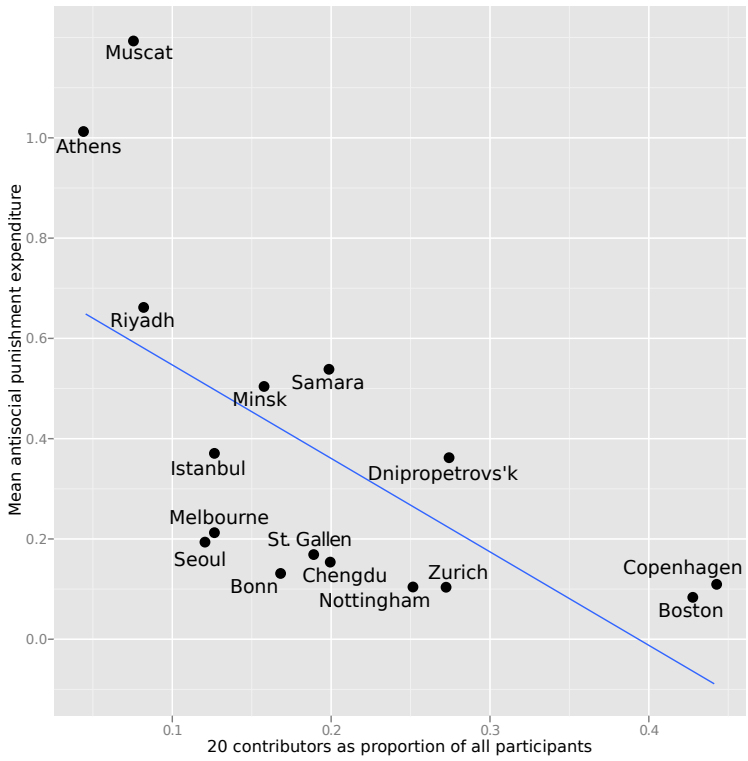


Fig. 4: Subject pools plotted by mean amount of ASP (y axis) and the proportion of subjects who contributed all of their available resources (20 tokens) in the first round.

4 Summary and Policy Implications

In the previous section we documented our contributions to the behavioural anthropology of human economic decision making, which we generated by taking an evolutionary approach. The assumption of this work is the standard one made in biology: that the seemingly bizarre behaviour of antisocial punishment 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 is 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 well-being 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 more likely to be expressed in contexts where group members do not expect by default that other group members are "in group". With respect to the previous point, this implies that there is always *some* cohort of trusted individuals, the question is how large it is by default. In Northern Europe (and Boston, the only US city surveyed here), it seems by default 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 *reduces ASP*. This implies that people inclined to ASP are impressed by such a clear expression of in-group assessment, and have some tendency to believe it and adopt it.

The final point may sound promising from a policy perspective. Note though that in the experimental context, commitment of resources is completely transparent. 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 make convincing.

Many of us can identify with the "in group" assessment that comes from knowing someone else has chosen the same college or university as we have, particularly in the same or similar year. An undergraduate degree is a significant investment — even where tuition is free, a degree requires 3–5 years of a person's life. For 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 is work for us. Most of us 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. The point is that in some cultures that feeling appears to extend even to the prestigious university campuses that Herrmann et al. (2008) chose to study[5]. This might indicate that it could be difficult 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 experimentation will identify in advance personality indicators for predisposition to ASP (e.g. Czibor and Bereczkei, 2012). On the other hand, 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.

4.1 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, as with many things, Hobbes was amazingly prescient concerning the creation of public goods given that he wrote in the

[5] 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.

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, yet generosity may in absence of other information be taken as an indication that in fact trust is merited.

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 well-being is (or at least appears to be) most determined by their relative dominance to the local neighbours, rather than to how well the neighbourhood performs as a whole, then it may be worth sacrificing income if longer-term benefit in terms of in-group status results. For organisations that *are* more concerned about global than local good, the best course of action is probably promoting the likelihood that the benefits of public goods are shared by those who are desired to cooperate, and ensuring transparency so all parties can be assured this is the case.

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 well being. In this case, the most useful avenue for future research would be to discover how easily or quickly the social characteristics leading to this failure can be altered. Measures available to be taken could have either cognitive (e.g. increased transparency in distribution of economic resources) or emotional (e.g. team building or other stage setting for triggering a state of emotional inclusiveness). If such measures work, a societies' citizens or leaders could be trained to recognise and exploit 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 development approaches.

Acknowledgements

We would like to thank Benedikt Herrmann for his help with theory building, education, citations, and his assistance in understanding his data set. Thanks also to Simon Gächter for meetings and occasional email assistance, and Karolina Sylwester 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.

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. (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.

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.

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.

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. (2008). Antisocial punishment across societies. *Science*, 319(5868):1362–1367.

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.

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

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.

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). Anti-social 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. (2011). *Homo homini lupus?* an evolutionary view on antisocial punishment. under review.

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

Szathmáry, E. (2011). To group or not to group? *Science*, 334(6063):1648–1649.

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.

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. including commentary and author's reply.